

**THE ROBUSTNESS OF IRT-BASED VERTICAL SCALING METHODS  
TO VIOLATION OF UNIDIMENSIONALITY**

by

**Liqun Yin**

B.A., East China Normal University, 1998

M.A. University of Pittsburgh, 2007

Submitted to the Graduate Faculty of  
School of Education in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH  
SCHOOL OF EDUCATION

This dissertation was presented

by

Liqun Yin

It was defended on

April 12, 2013

and approved by

Clement A. Stone, Ph.D., School of Education

Feifei Ye, Ph.D., School of Education

Levent Kirisci, Ph.D., School of Pharmacy

Dissertation Advisor: Suzanne Lane, Ph.D., School of Education

Copyright © by Liquun Yin

2013

# **THE ROBUSTNESS OF IRT-BASED VERTICAL SCALING METHODS TO VIOLATION OF UNIDIMENSIONALITY**

Liqun Yin, PhD

University of Pittsburgh, 2013

In recent years, many states have adopted Item Response Theory (IRT) based vertically scaled tests due to their compelling features in a growth-based accountability context. However, selection of a practical and effective calibration/scaling method and proper understanding of issues with possible multidimensionality in the test data is critical to ensure their accuracy and reliability. This study aims to use Monte Carlo simulation to investigate the robustness of various unidimensional scaling methods under different test conditions and different degrees of departure from unidimensionality in common-items nonequivalent groups design (grades 3 to 8). The main research questions answered by this research are: 1) Which calibration/scaling methods, concurrent, semi-concurrent, separate calibration with SL scaling, separate calibration with mean/sigma scaling, and pair-wise calibration, yield least biased ability estimates in the vertical scaling context? 2) How do different degrees of multidimensionality affect use of the methods?

Results indicate that various calibration and scaling methods perform very differently under different test conditions, especially when the grades are furthest away from the base grade. Under unidimensional condition, the five calibration and linking methods produced very similar results when the grades are close to the base grade 5. However, for grades 7 and 8, semi-concurrent and concurrent calibrations yielded more biased results while the results for the other three are comparable. Under multidimensional conditions, all five methods produced more biased results and the bias

patterns differed across methods. In general, the more severe the multidimensionality is, the larger the biases are. Among the five methods compared, separate calibration with SL linking is the most robust to variations in multidimensionality.

## TABLE OF CONTENTS

<b>PREFACE .....</b>	<b>XIV</b>
<b>1.0 INTRODUCTION .....</b>	<b>1</b>
<b>1.1 STATE OF THE PROBLEM.....</b>	<b>1</b>
<b>1.1.1 Background.....</b>	<b>1</b>
<b>1.1.2 Attractive features of IRT-based vertical scaling .....</b>	<b>3</b>
<b>1.1.3 Issues related to constructing an IRT-based vertical scale .....</b>	<b>4</b>
<b>1.1.4 Calibration and Scaling Methods .....</b>	<b>5</b>
<b>1.1.5 Summaries .....</b>	<b>7</b>
<b>1.2 PURPOSE .....</b>	<b>8</b>
<b>1.3 RESEARCH QUESTIONS .....</b>	<b>8</b>
<b>1.4 SIGNIFICANCE OF THE STUDY.....</b>	<b>9</b>
<b>2.0 LITERATURE REVIEW .....</b>	<b>11</b>
<b>2.1 DATA COLLECTION DESIGNS.....</b>	<b>11</b>
<b>2.2 UNIDIMENSIONAL IRT MODELS.....</b>	<b>14</b>
<b>2.3 CALIBRATION AND SCALING METHODS.....</b>	<b>17</b>
<b>2.3.1 Scale indeterminacy .....</b>	<b>17</b>
<b>2.3.2 Concurrent calibrations and separate calibrations .....</b>	<b>19</b>
<b>2.3.3 Scale transformation methods for unidimensional model.....</b>	<b>22</b>
<b>2.3.4 Moments methods .....</b>	<b>23</b>

2.3.5	Characteristic curve methods .....	25
2.3.6	Minimum $\chi^2$ method and least squares method .....	26
2.3.6.1	Proficiency transformation method .....	27
2.3.6.2	Comparison studies of scale transformation methods .....	28
2.3.7	Comments about choosing scale transformation methods .....	33
2.4	ABILITY ESTIMATION .....	35
2.4.1	Maximum likelihood estimation (MLE).....	35
2.4.2	Maximum a posterior (MAP) and Expected a posterior (EAP) ..	36
2.4.3	Comparisons of ability estimation methods.....	37
2.5	FACTORS RELATED TO VERTICAL SCALING .....	41
2.5.1	Number of grade levels and choice of a base year.....	41
2.5.2	Sample size .....	43
2.5.3	Test length and number of common items.....	43
2.6	DIMENSIONALITY ISSUES.....	44
2.6.1	Unidimensionality assumption.....	44
2.6.2	Multidimensional data structures.....	46
2.6.3	Effects of multidimensionality on IRT-based vertical scaling .....	48
2.7	SUMMARY.....	51
3.0	METHODS.....	53
3.1	DESIGN OF THE SIMULATION STUDY .....	53
3.1.1	Vertical scaling data collection design.....	53
3.1.2	IRT model .....	56
3.1.3	Manipulated factors .....	58
3.1.3.1	Intertrait Correlation .....	58

3.1.3.2	Calibration and scaling methods.....	60
3.1.3.3	The variances of ability distributions for 6 grades.....	67
3.1.3.4	Fixed factors .....	68
3.1.4	Number of replications .....	69
3.2	EVALUATION AND COMPARISON CRITERIA .....	70
3.3	DATA GENERATION .....	71
3.3.1	Software for data generation, calibrations, and scaling .....	71
3.3.2	Data generation .....	72
3.3.3	Validating data generation .....	76
3.3.4	Implementation of the study design.....	78
4.0	RESULTS .....	83
4.1	COMPARISONS OF SCALING METHODS UNDER UNIDIMENSIONAL CONDITION.....	84
4.1.1	Bias of ability estimates .....	84
4.1.2	RMSD of ability estimates .....	89
4.1.3	Correlation between estimated and “true” abilities .....	94
4.1.4	Summary .....	97
4.2	COMPARISONS OF SCALING METHODS UNDER MULTIDIMENSIONAL CONDITON WITH $R = .3$ BETWEEN TWO LATENT ABILITIES.....	100
4.2.1	Bias of ability estimates .....	100
4.2.2	RMSD of ability estimates .....	105
4.2.3	Correlation between estimated and “true” abilities .....	112
4.2.4	Summary .....	115



4.3	COMPARISONS OF SCALING METHODS UNDER MULTIDIMENSIONAL CONDITON WITH $R = .6$ BETWEEN TWO LATENT ABILITIES.....	116
4.3.1	Bias of ability estimates .....	116
4.3.2	RMSD of ability estimates .....	120
4.3.3	Correlation between estimated and “true” abilities .....	124
4.3.4	Summary .....	126
5.0	DISCUSSION.....	130
5.1	MAJOR FINDINGS OF THIS STUDY .....	130
5.1.1	Answers to research questions .....	131
5.1.2	Summary of major findings .....	135
5.1.3	Implications to educators and research practitioners .....	141
5.2	LIMITATIONS AND FUTURE STUDY.....	143
	APPENDIX A SELECTED MPLUS RESULTS .....	145
	APPENDIX B SAS PROGRAM TO SIMULATE VERTICAL SCALING.....	149
	APPENDIX C SELECTED MULTILOG FILES .....	155
	BIBLIOGRAPHY.....	160

## LIST OF TABLES

Table 4.1 Average bias of $\theta$ under unidimensional condition with fixed variance.....	87
Table 4.2 Average bias of $\theta$ under unidimensional condition with varied variances .....	87
Table 4.3 Average RMSDs of $\theta$ under unidimensional condition with fixed variance ..	92
Table 4.4 Average RMSDs of $\theta$ under unidimensional condition with varied variances	92
Table 4.5 Average correlation of $\theta$ and $\theta$ under unidimensional condition with fixed variance.....	96
Table 4.6 Average correlation of $\theta$ and $\theta$ under unidimensional condition with varied variances .....	96
Table 4.7 Average bias of $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance.....	103
Table 4.8 Average bias of $\theta$ under multidimensional condition ( $r=0.3$ ) with varied variances .....	103
Table 4.9 Average RMSD of $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance.....	110
Table 4.10 Average RMSD of $\theta$ under multidimensional condition ( $r=0.3$ ) with varied variances .....	110
Table 4.11 Average correlation of $\theta$ and $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance .....	114

Table 4.12 Average correlation of $\theta$ and $\theta$ under multidimensional condition( $r=0.3$ ) with varied variances .....	114
Table 4.13 Average bias of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	118
Table 4.14 Average bias of $\theta$ under multidimensional condition ( $r=0.6$ ) with varied variances .....	118
Table 4.15 Average RMSD of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	122
Table 4.16 Average RMSD of $\theta$ under multidimensional condition ( $r=0.6$ ) with varied variances .....	122
Table 4.17 Average correlation of $\theta$ and $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance .....	125
Table 4.18 Average correlation of $\theta$ and $\theta$ under multidimensional condition( $r=0.6$ ) with varied variances .....	125

## LIST OF FIGURES

Figure 2.1 Multidimensional data structure.....	47
Figure 3.1 Common-items non-equivalent groups design. ....	55
Figure 3.2 Graphical representation of the separate, semi-concurrent, pair-wise, and concurrent calibrations .....	67
Figure 3.3 Simulated multidimensional data structure.....	75
Figure 3.4 Flow chart of the simulation study implementation.....	82
Figure 4.1 Average bias of $\theta$ under unidimensional condition with fixed variance .....	88
Figure 4.2 Average bias of $\theta$ at unidimensional condition with varied variances .....	88
Figure 4.3 Average RMSDs of $\theta$ under unidimensional condition with fixed variance.	93
Figure 4.4 Average RMSDs of $\theta$ under unidimensional condition with varied variances .....	93
Figure 4.5 Average bias of $\theta$ at multidimensional condition ( $r=0.3$ ) with fixed variance .....	104
Figure 4.6 Average bias of $\theta$ under multidimensional condition ( $r=0.3$ ) with varied variances.....	104
Figure 4.7 Average RMSD of $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance.....	111

Figure 4.8 Average RMSD of $\theta$ under multidimensional condition ( $r=0.3$ ) with varied variances.....	111
Figure 4.9 Average bias of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	119
Figure 4.10 Average bias of $\theta$ under multidimensional condition ( $r=0.6$ ) with varied variances.....	119
Figure 4.11 Average RMSD of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	123
Figure 4.12 Average RMSD of $\theta$ under multidimensional condition ( $r=0.6$ ) with varied variances.....	123
Figure 4.13 Average bias of $\theta$ under unidimensional condition with fixed variance ...	127
Figure 4.14 Average bias of $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance.....	127
Figure 4.15 Average bias of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	127
Figure 4.16 Average RMSDs of $\theta$ under unidimensional condition with fixed variance .....	128
Figure 4.17 Average RMSDs of $\theta$ under multidimensional condition ( $r=0.3$ ) with fixed variance.....	128
Figure 4.18 Average RMSDs of $\theta$ under multidimensional condition ( $r=0.6$ ) with fixed variance.....	128

## **PREFACE**

This dissertation would not have been completed without the help and support from many people, to whom I'd like to express my sincere gratitude here.

First and foremost, I want to thank Dr. Suzanne Lane, a tremendous mentor. Her guidance, care, and support throughout my graduate study have played an instrumental role during this journey. Her great insights and guidance helped in every stage of my dissertation. Before I started in the program, I met with Dr. Lane with various concerns that posed to be an uphill battle. I will never forget what she said to me at that first meeting, "Everything will be fine." That simple sentence gave me much ease and I walked away feeling confident in my decision to pursue a doctorate that day. I will forever cherish the memory of being her student.

Secondly, I would like to thank Dr. Clement Stone for spending much of his valuable time answering my questions and co-supervising my dissertation. His many thoughtful comments and suggestions helped to make the dissertation both stronger and better.

I am also grateful to Dr. Feifei Ye and Dr. Levent Kirisci for serving on my dissertation committee. Their constructive input and guidance further enhanced the quality of this study. I would also like to thank Dr. Kevin Kim for his recommendation

and supervision and for giving me the opportunity to apply my knowledge to real world scenarios.

Last but not least, I would like to thank my family. Thank you from the bottom of my heart to my parents for their continued encouragement and support. Since my undergraduate days, they have stood behind me in my pursuit of being a teacher and have always instilled a love of learning in me.

I would also like to acknowledge my two wonderful sons, Michael and Jared. You both make me feel happy and proud each day. I hope through my example that they will be inspired to dream big, to never stop learning and to never give up. I am hopeful for their future and wish that it may always be filled with love, joy and laughter.

At the end of it all, I could not have accomplished this goal without the unconditional love of my beloved husband, Dr. Minhong Mi. The immense sacrifices he made during this time coupled with his unfailing love and support has allowed me to not only pursue but also achieve a lifelong dream.

## **1.0 INTRODUCTION**

### **1.1 STATE OF THE PROBLEM**

#### **1.1.1 Background**

Over the past several decades, test-based accountability systems have played a prominent role in educational reforms (Linn, 2000). The recent federal act, *No Child Left Behind Act of 2001* (NCLB, 2001), further strengthens the trend by using test-based accountability systems as educational reform tools (Patz, 2007; Harris, 2007). However, determining the most effective way to implement test-based accountability in practice is challenging for practitioners, policy makers, and educational researchers. One of the most difficult challenges is to find appropriate ways to measure grade-to-grade growth or academic change in student achievement (Kolen & Brennan, 2004; Lissitz & Huynh, 2003).

Two statistical models, growth models and value-added models (VAM), are being used for evaluating students' growth and schooling effects under NCLB accountability systems (Patz, 2007; Haertel, 2005; Schmidt, Houang, & McKnight, 2005). VAM uses the annual standardized test scores for individual students, usually administered at the end of the school year, to measure their progress in core academic subjects, and applies the results as a measure of both the non-schooling effects and schooling effects such as teacher effects and school effects (Martineau, 2006; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Stewart, B. E., 2006). Growth



models also use longitudinal data to measure students' growth or change either at the individual level or at the school level (Briggs & Weeks, 2009; Haertel, 2005). The simplest use of growth models is to summarize and investigate the trends in students' achievement across academic years such as the reading ability from grade 3 through 8. In most cases, these trends are compared to a criterion to evaluate how much growth is considered as adequate or inadequate (Haertel, 2005; Briggs & Weeks, 2009; McCall, 2007).

Although there are some differences between growth models and VAMs, one of the central assumptions underlying these two methods is that different test scores can be vertically scaled to a common scale so that they are comparable over time (Briggs, Weeks, & Wiley, 2008). Vertical scaling (also referred to as vertical linking) is a statistical method that attempts to put test scores on a common scale for tests that measure the same general domain of skills or constructs, but are intentionally designed to be different in content and difficulty (Kolen & Brennan, 2004).

The goal of vertical scaling is to position the tests and groups of examinees along a common scale (Baker, 1984). After vertical scaling, test scores are considered comparable over time or across grades (Kolen & Brennan, 2004; Loyd & Hoover, 1980). More importantly, with a vertical scale, the students' or schools' year-to-year growth can be demonstrated by the change of scale scores. Therefore, vertical scaling plays a significant role in shaping adequate yearly progress and growth-based accountability programs. In recent years, state assessment programs, such as those in North Carolina, California, Colorado, Tennessee, Oregon, Idaho, and Florida, have begun to use vertically scaled tests as part of their accountability programs.

### **1.1.2 Attractive features of IRT-based vertical scaling**

Two types of methods are commonly used for developing a vertical scale, the Thurstonian methods and the IRT-based methods (Yen & Burket, 1997; Kolen & Brennan, 2004; Yen, 1986; Williams, Pommerich, & Thissen, 1998).

The Thurstonian methods are traditional vertical scaling methods that use summed scores for scaling. The basic idea of the Thurstonian methods is that there is a linear relationship between the z-scores associated with the distribution of two summed scores. However, the scaled scores are dependent upon not only the ability of examinees but also other characteristics of the population from which the groups are drawn, and even the interaction of ability and the context of sampled items. Therefore, the linear relationship of two sets of z-scores will not always hold for all groups if there is sample dependency. In response to the complications created by sample dependency of Thurstonian scaling methods, IRT-based vertical scaling methods received attention in the early 1980s (Loyd & Hoover, 1980; Yen & Burket, 1997).

IRT-based scaling places all items on the same scale. As a result, it possesses several attractive characteristics (Harris & Hoover, 1987; Kolen & Brennan, 2004; Skaggs & Lissitz, 1986; Karkee, Lewis, Hoskens, Yao, & Haug, 2003): (1) It takes not only the student's latent ability, but also item characteristics, into account for estimating student proficiency. (2) It is continuous with equal intervals, which is close to the nature of developmental year-to-year growth. (3) When data fits the model well, it will demonstrate the desired invariance properties, *i.e.*, "test-free" and "sample-free" latent ability and item parameter estimates. For these reasons, IRT-based scaling methods are the focus of this study.

### **1.1.3 Issues related to constructing an IRT-based vertical scale**

As compared to the Thurstonian methods, IRT-based vertical scaling makes stronger assumptions. In particular, it assumes the items to be scaled are unidimensional or essentially unidimensional (Kolen & Brennan, 2004). That is, item responses are based on the same skill or same composite of multiple skills (Walker, Azen, & Schmitt, 2006; Ackerman, 1994).

According to several researchers (Schmidt, Houang, & Mcknight, 2005; Lockwood, *et al.*, 2007; Martineau, 2006), the unidimensional assumption is very likely to be violated in longitudinal cross-grades tests. Instead, the tests are likely to exhibit a multidimensional structure across grades. Recently, some researchers proposed to use multidimensional IRT (MIRT) to construct vertical scales (Patz & Yao, 2007; Yao & Mao, 2004; Reckase & Martineau, 2004), so as to increase the accuracy of vertical scaling. Obviously, the MIRT-based vertical scaling method is much more complicated and is still in an early development stage.

Most educational and psychological tests are still using unidimensional IRT-based scaling methods to construct a common scale even if the tests are likely to exhibit some degree of multidimensional structure across grades (Reckase & Martineau, 2004; Lissitz, & Huynh, 2003; Reckase, 1985). However, very limited research has been conducted on the robustness of unidimensional IRT scaling methods in the vertical scaling context. It remains an open question whether a unidimensional IRT model can be successfully applied to a vertical scale that may not be strictly unidimensional in nature (Boughton, Lorie, & Yao, 2005).

Therefore, investigating the robustness of scaling methods to the multidimensionality condition in the vertical scaling context is the main motivation of this study. More specifically, this study tries to determine the robustness of unidimensional vertical scaling methods for the multidimensionality data across grade levels.

#### **1.1.4 Calibration and Scaling Methods**

In IRT-based vertical scaling, to estimate examinees' latent ability or proficiency, item parameters need to be estimated and put on a common scale. The process of estimating the item parameters is referred to as "calibration". There are two general IRT-based calibration methods in vertical scaling: concurrent and separate calibrations (Hanson & Beguin, 2002; Kolen & Brennan, 2004).

With the concurrent calibration method, a vertical scale is established by calibrating item responses from all grade levels in a single computer run. After the concurrent calibration is conducted, the item parameter estimates for all grades are on the same scale.

With the separate calibration method, the estimations are usually based on different scales due to the IRT scale indeterminacy problem (Lord, 1980). To put item parameters on a common metric, a set of scale transformations must be performed based on the information from common items. The procedures for finding transformation parameters and putting the estimates from separate calibrations on a common scale is referred to as "scaling", "transformation", or "linking" (Kolen & Brennan, 2004; Kim & Cohen, 1998).

There are several popular scaling methods: Mean/mean method (Loyd & Hoover, 1980), Mean/sigma method (Marco, 1977), item characteristic curves method (ICC method, Haebara, 1980), and test characteristic curves method (TCC method, Stocking & Lord, 1983). The first two are moments methods, which are attractive because of their statistical simplicity. The latter two are characteristic curves methods, which use more available information from item parameters and thus are expected to produce more adequate scaling results (Baker & Al-Karni, 1991; Hanson & Béguin, 2002).

There is some controversy in the adequacy among the calibration and scaling methods (Kolen & Brennan, 2004; Briggs & Weeks, 2009; Karkee, Lewis, Hoskens, & Yao, 2003). Concurrent calibration is simple to implement and might be preferred in horizontal equating (Hanson & Béguin, 2002). However, the robustness to violations of unidimensionality is questionable under vertical scaling. Separate calibrations are thought to be safer in vertical scaling since the unidimensionality assumption is likely to be violated across multiple grades. However, the long-chained scaling procedures involved in separate calibrations might introduce more measurement errors and are more time consuming than concurrent estimation methods. Additionally, different linking methods will likely produce different scaling results. Recently, a new hybrid calibration, semi-concurrent calibration, has been proposed (Meng, 2007; Briggs & Weeks, 2009; Karkee, Lewis, Hoskens, & Yao, 2003). The hybrid calibration combines the strengths of concurrent and separate calibrations for estimation and is expected to produce more adequate scaling results.

Many studies have focused on comparisons of different calibration methods under the horizontal equating framework (Hanson & Béguin, 2002; Kim & Cohen, 1998; Kim & Cohen, 2002; Baker & Al-Karni, 1991; Lee & Ban, 2010; Hu, Rogers, &

Vukmirovic, 2008). A limited number of comparison studies have been conducted under the vertical scaling context (Briggs & Weeks, 2009; Ito, Sykes, & Yao, 2008; Karkee et al., 2003; Meng, 2007; 2003). In these studies, the TCC scaling method is the only transformation method used for separate calibrations. In addition, the conclusions about the behaviors of different calibration methods are not consistent across the studies. These discrepancies suggest that more comprehensive comparisons among different calibration and scaling methods are needed under different conditions in the vertical scaling context.

#### **1.1.5 Summaries**

Vertical scaling facilitates the ability estimation and tracking of students' growth using repeated measures on individuals across years. Both psychometric factors and practical problems affect the accuracy of IRT-based vertical scaling results. The main psychometric factors include the selection of practical vertical scaling designs, number of common items, and various vertical scaling methods. The major practical issue is the inevitable violation of the unidimensionality assumption while developing a vertical scale across multiple grades. Comparing various IRT-based scaling methods and selecting the most appropriate one is very important in vertical scaling, especially under accountability systems, since potentially inaccurate results will affect the validity of the score interpretation and distort conclusions about the test results. Furthermore, very limited research has been done investigating the robustness of different vertical scaling methods. Therefore, to provide more information about the behaviors of different scaling

methods, different calibration and scaling methods need to be compared for different test conditions.

## **1.2 PURPOSE**

The main purpose of this dissertation is to use Monte Carlo studies to investigate the robustness of various unidimensional calibration and scaling procedures to the inevitable violation of the unidimensional assumption while developing vertical scales. Using common-items nonequivalent groups as the vertical scaling design, the different calibration and scaling methods are compared under both unidimensionality and non-unidimensionality conditions with different ability variance structures.

## **1.3 RESEARCH QUESTIONS**

Three main research questions are addressed in this dissertation.

1) When the IRT unidimensionality assumption holds, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma (MS), and separate with SL(or TCC) linking, affect the vertical scaling results with assumed different latent ability variance structures for a wide range of grades?

2) When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is low, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise

calibration, separate with mean/sigma (MS), and separate with SL (or TCC) linking, affect the vertical scaling results with assumed different latent ability variance structures for a wide range of grades?

3) When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is moderate, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma (MS), and separate with SL (or TCC) linking, affect the vertical scaling results with assumed different latent ability variance structures for a wide range of grades?

#### **1.4 SIGNIFICANCE OF THE STUDY**

State assessment programs often vertically scale test results across grade levels, with the common-item non-equivalent groups design being the most popular vertical scaling design. Because there is limited research on the vertical scaling design, especially on the investigation of robustness of various vertical scaling methods, the results reported in this study have significant implications for applied researchers and testing professional at state assessment programs. Since assessment programs often vertically scale results across grade levels, it is critical that practitioners select an effective and practical vertical scaling method. It is also important for testing professionals to understand how the scaling results are affected by many technical factors and practical issues, including the use of different calibration and scaling methods under various degrees of violation of the unidimensionality assumption.



This study investigates the behaviors of different calibration and scaling methods under the common-item non-equivalent groups design. More specifically, it investigates the robustness of various unidimensional calibration and scaling procedures to the violation of the unidimensional assumptions while developing vertical scales. It not only addresses several questions with regard to the adequacy of various vertical scaling methods but also provides practical guidance for selecting an effective vertical scaling method under different test conditions for practitioners and testing professionals at state assessment programs.

## **2.0 LITERATURE REVIEW**

In this chapter, the literature on IRT-based vertical scaling is reviewed. More specifically, vertical scaling data collection designs are illustrated. Next, the main technical procedures in vertical scaling are introduced. It includes a discussion on the nature of IRT models, calibration and scaling methods, and proficiency score estimation methods. Then, some practical factors related to IRT vertical scale are discussed, with a focus on the selection of a base year, the number of common items, and on the violation of IRT assumptions. Finally, research on vertical scaling is reviewed to provide a rationale for the selection of the investigated control factors in this dissertation.

### **2.1 DATA COLLECTION DESIGNS**

Three vertical scaling data collection designs can be used to develop a vertical scale: common-item nonequivalent groups design (sometimes also called a common-item design), equivalent groups design, and scaling test design (Kolen & Brennan, 2004; McCall, 2007; McBride & Wise, 2001).

In a scaling test design, each examinee is administered the test level appropriate for her or his grade. This test is also called a grade level test. In addition, a scaling test is

administered to students across grades. The scaling test is a special test that spans the content across all of the grade levels such as grades 3 to 8. Students in all of the grades are administered the same scaling test. The scaling test is used as external common items for all students to establish a vertical scale. However, constructing a scaling test to represent content areas across all grade levels is very difficult and impractical for many assessment programs. This design is also very likely to have floor effects and ceiling effects since a scaling test can be extraordinarily difficult for lower grade level students and very easy for higher grade level students. These undesired testing effects will further invalidate the scaling results. Therefore, the use of a scaling test design is limited in practice.

In an equivalent groups design, the students are randomly selected to take either their grade level test or the adjacent grade level test (Kolen & Brennan, 2004). The two groups of students are considered as equivalent groups. The vertical scale is constructed through the information from the equivalent groups, or say, common examinees. The limitation of this design is that the validity of the test results may be questionable because the assumed equivalent groups, not the same examinees, take the tests with different difficulty.

Among three vertical scaling designs, the common-item design is the easiest and most widely used vertical scaling design. In the common-item design, tests appropriate for each grade level (referred to as level tests) are constructed with a set of common item blocks. Common items are the items overlapping between any pair of adjacent grades.

The common items could be only from the lower grade level of the adjacent grades, or from the higher grade level of the adjacent grades, or from both of the two adjacent grades. According to the different types of common item blocks, three specific

linking designs are categorized, below-grade/on-grade design, on-grade/above-grade design, and below-grade/on-grade/above-grade design, respectively (Wang, Jiao, Young, & Jin, 2005). Among them, below-grade/on-grade design may be more preferable than the other two because it minimizes random guessing effects. If students take the items in a higher grade level, it is more likely that randomly guessing will occur.

When common items are used, the differences in ability levels of the groups are reflected in the values of the common item difficulty estimates, and the differences in group variability are embedded in the item discrimination values yielded by the common items (Baker, 1984). These common item parameter estimates are then used to place item parameter estimates from each grade onto a common scale. The common scale is not unique and is affected by many factors such as the vertical scaling design and scaling methods. These factors and issues are discussed in the following sections.

Although the common-item design is the most widely used vertical scaling design, it still has several limitations, one of which is that it uses out-of-level items to construct the common item block. In the common-item vertical scaling design, the content and difficulty of common items usually span multiple grades. If these common items are on-level items for one grade, they could be out-of-level items for the adjacent grades, either easier or more difficult for the adjacent grade students. If students are given items that are too difficult or too easy for them, the resulting data are may be poor quality with associated “floor” and “ceiling” effects (Haertel, 2005). In addition, using tests of inappropriate difficulty typically leads to large conditional standard error of measurements (CSEM). More importantly, the associated measurement errors are considered as another artifactual cause of scale shrinkage (Camilli, Yamamoto, & Wang,

1993). Therefore, the common-item design itself has limitations for developing a vertical scale.

Another limitation of the common-item design is that it is subject to context effects (Kolen & Brennan, 2004). A context effect occurs when an examinee's item responding behavior is affected by the location of an item within a test (Kingston & Dorans, 1984). In other words, common items when placed at different positions in tests most likely will behave differently. This sort of context effect threatens item parameter invariance and creates systematic error in the scaling or linking (Meyers, Miller, & Way, 2009). For horizontal equating, one possible solution is to keep the common items in similar positions for the two test forms. However, this method can be very limiting and difficult to sustain in vertical scaling since the item difficulties increase across grade levels. Another possible solution to context effects is the elimination of items that are not sufficiently resistant to location effects (Kingston & Dorans, 1984; Meyers, *et. al.*, 2009).

## **2.2 UNIDIMENSIONAL IRT MODELS**

A variety of IRT models have been developed. Based on measured dimensions, IRT models can be divided into unidimensional models and multidimensional models. Based on item types, IRT models can be divided into dichotomous and polytomous IRT models.

In unidimensional IRT models, a single latent trait is assumed to characterize person differences. A unidimensional IRT model is appropriate for data in which a

single common factor underlies item responses. It also can be used for data in which two or more factors underlie item responses if these factors have similar combinations across all items (Embretson & Reise, 2000). If data measure two or more factors and these factors have differing impact on item responses, multidimensional IRT models are preferred.

For binary data, item responses are scored for only two outcomes, correct versus incorrect. Both unidimensional IRT models and multidimensional IRT models are available for binary data. There are two types of unidimensional IRT models for binary data. One is the traditional logistic model and the other is the normal ogive model. Logistic models are the most popular unidimensional models for binary data due to their computationally simple form. Based on the number of item parameters used in the model, logistic models may be categorized as the one-parameter logistic model (1P or Rasch model), the two-parameter logistic model (2P model), and the three-parameter logistic model (3P model).

The 3P logistic IRT model is based on the logistic distribution, which gives the probability of a response as follows:

$$P_i(\theta|a, b, c) = c + (1 - c) \frac{e^{Da(\theta-b)}}{1 + e^{Da(\theta-b)}} \quad (1)$$

where  $\theta$  is the proficiency or latent ability level of an examinee;  $e$  is the natural log base ( $e = 2.718$ );  $a$  is the item discrimination parameter or the slope of the item response curve. The item discrimination parameter usually varies between .5 and 2.0. The higher an  $a$  value, the more discriminating the item is among the examinees.  $b$  is the item difficulty parameter or location parameter of the item response curve.  $c$  is the

guessing parameter or lower asymptote for the item. It specifies the probability of a correct response for examinees who have very low ability levels. The value of  $c$  is usually between 0 and 0.3 for multiple choice items.  $D$  is a constant and is equal to 1.702.  $P_i(\theta|a, b, c)$  is the probability of an examinee with proficiency score  $\theta$  responding correctly to an item. Since three item parameters,  $a$ ,  $b$ , and  $c$ , are used in the model, this model is usually called three-parameter (3P) logistic IRT model.

In the above logistic model, if the guessing chance is very low, the item parameter  $c$  could be set at 0 for all items. Then the probability of correct response for the examinees only relates to the  $a$  and  $b$  item parameters. This model is called a 2P logistic IRT model.

Further, if the item discrimination parameter,  $a$ , is assumed to be constant such as 1 for all items, the probability of correct response for the examinees can be expressed as following:

$$P_i(\theta|b) = \frac{e^{D(\theta-b)}}{1+e^{D(\theta-b)}} \quad (2)$$

Equation 2 shows that the probability of a correct response for the examinees is only dependent on the item difficulty parameter and examinee's ability. This model is called the 1P logistic IRT model or Rasch model (Lord, 1980). The Rasch model and 3P logistic model have been the most prominent IRT models used in horizontal equating and vertical scaling. The Rasch model possesses certain desirable properties such as statistical simplicity, the practical elimination of guessing, statistical efficiency for small sample size, and a monotonic relationship between raw scores and the estimated latent

ability of examinees (Harris & Hoover, 1987, Loyd & Hoover, 1980; Skaggs & Lissitz, 1988).

Although the Rasch model shows a great deal of promise for improving the horizontal equating of tests, the Rasch model appears not to work satisfactorily for vertical scaling of multiple-choice tests (Holmes, 1982; Loyd & Hoover, 1980; Slinde & Linn, 1978). In addition, the Rasch model is less robust to violations of its assumptions in vertical scaling than for horizontal equating (Skaggs & Lissitz, 1986). In the vertical scaling context, because the item difficulty parameters usually are across a wide range of grades, the assumptions of no guessing and uniform discriminations are more likely violated than in the horizontal equating context. For vertical scaling, the 3P logistic model has generally demonstrated better results than the Rasch model (Skaggs & Lissitz, 1988; Harris & Hoover, 1987). Therefore, the 3P logistic IRT model is the selected IRT model in this simulation study.

## **2.3 CALIBRATION AND SCALING METHODS**

### **2.3.1 Scale indeterminacy**

Based on item response theory, the origin and unit of the measurement of the latent ability metric are undetermined and usually defined arbitrarily to some extent. This characteristic is often referred to as the IRT scale indeterminacy problem (Lord, 1980). More specifically, the scale used for the proficiency score ( $\theta$ ) is, in theory, determined by an arbitrary linear transformation.



Using the 3P logistic unidimensional IRT model as an example (Stocking & Lord, 1983), the probability of an examinee with proficiency score  $\theta$  responding correctly to an item,  $P_i(\theta|a, b, c)$  is a function of  $a(\theta - b)$ .

Suppose that  $\theta$ ,  $a$ ,  $b$ , and  $c$  are transformed by a series of linear transformation to a given  $\theta^*$ ,  $a^*$ ,  $b^*$ ,  $c^*$ , and

$$\theta^* = A\theta + B \quad (3)$$

$$b^* = Ab + B \quad (4)$$

$$a^* = \frac{1}{A}a \quad (5)$$

$$c^* = c \quad (6)$$

where  $A$  is the slope of the linear transformation;  $B$  is the intercept of the linear transformation;  $c$  is not changed. According to the 3P logistic IRT model, the  $P_i(\theta^*)$  is a function of  $a^*(\theta^* - b^*)$ . From equations 3, 4, and 5,

$$a^*(\theta^* - b^*) = \frac{1}{A}a [(A\theta + B) - (Ab + B)] = a(\theta - b) \quad (7)$$

Since  $c^* = c$ , one can obtain  $P_i(\theta^*|a^*, b^*, c^*) = P_i(\theta|a, b, c)$  from equation (7). In other words, the probability of an examinee correctly responding to an item will not be affected by the ability scale transformation when the item parameters are transformed at the same time. Therefore, the problem of transforming the scales in IRT-based equating and scaling reduces to the problem of finding the appropriate linear transformation parameters  $A$  and  $B$ , to obtain adequate equating results. In the common

items nonequivalent groups design, the information from common items is usually used to obtain the estimates of the scale transformation parameters,  $A$  and  $B$ . Once  $A$  and  $B$  are obtained, the values of  $A$  and  $B$  are substituted into equations (3), (4), and (5) to get the rescaled parameters estimates,  $\theta^*$ ,  $a^*$ , and  $b^*$ . After these transformations, the estimates from two separate calibrations are now on the same scale.

### **2.3.2 Concurrent calibrations and separate calibrations**

In IRT, the process of estimating item parameters is referred to as calibration (Kolen & Brennan, 2004; Meng, 2007). When conducting equating or scaling with nonequivalent groups, the parameters from different test forms are usually on different scales and need to be put on a common scale.

Traditionally, there have been two IRT approaches used to create a common vertical scale across two or more ability levels: concurrent and separate calibration (Hanson & Beguin, 2002; Kolen & Brennan, 2004). In concurrent calibration, a common scale is established by calibrating data from all grade levels in a single computer run. That is, an examinee's item response pattern contains item responses for the items taken and a "not reached" code for the items not taken (Kolen & Brennan, 2004). After concurrent calibration, the item parameter estimates and ability estimates for all grades are already on the same scale. Scale transformations are not necessary under this procedure.

When a separate calibration is used, the IRT parameters are estimated separately for each grade. As discussed above, there is a scale determinacy problem for IRT logistic models. To solve this problem, the scale used for the proficiency score ( $\theta$ ) is

usually defined as a  $N(0,1)$  for each grade by computer default. Therefore, if separate calibrations are conducted, the item parameter estimates are usually based on different scales. To put item parameters on a common metric, a set of scale transformations must be performed based on the information from common items. The procedures of finding transformation parameters and putting the estimates from separate calibrations on a common scale are referred to as a “scaling method” or a “linking method” (Kolen & Brennan, 2004; Kim & Cohen, 1998).

There is controversy in the adequacy among the calibration methods (Kolen & Brennan, 2004; Briggs & Weeks, 2009; Karkee, Lewis, Hoskens, & Yao, 2003). Concurrent calibration is simple to implement and might be preferred in horizontal equating (Hanson & Beguin, 2002), but its robustness to violations of unidimensionality is questionable in vertical scaling. The separate calibrations are thought to be better for vertical scaling since the unidimensionality assumption across multiple grades is likely to be violated. However, the different linking methods involved in separate estimations might introduce more measurement error and separate estimations are more time consuming than concurrent estimation. In addition, different linking methods likely produce different scaling results.

Recently, a new hybrid calibration (also called a modified concurrent calibration) has been proposed by researchers (Meng, 2007; Briggs & Weeks, 2009; Karkee, Lewis, Hoskens, & Yao, 2003; Ito, Sykes, & Yao, 2008). The hybrid calibration is a combination of concurrent and separate calibrations. It was originally proposed to address the problems introduced by multidimensional data. There are three specific hybrid calibration methods: pair-wise concurrent calibration (Karkee, Lewis, Hoskens, & Yao, 2003; Briggs & Week, 2009), semi-concurrent calibration (Meng, 2007), and

separate grade-groups calibration (Ito, Sykes, & Yao, 2008). In a pair-wise concurrent calibration, each non-overlapping pair of adjacent grades are estimated simultaneously, and a common scale is then constructed by using scale transformation methods. In a semi-concurrent calibration, each half of the grade levels are estimated simultaneously. A set of scale transformations are then used to put the two scales on a common scale. Similarly, in separate grade-groups calibration, a few of the adjacent grades are calibrated simultaneously. Then a set of transformations are used to put several calibrations on a common scale. Theoretically, the hybrid calibration combines the strengths of concurrent and separate calibrations and is expected to produce more adequate scaling results.

Many studies have focused on comparisons among different calibration methods under the horizontal equating framework (Hanson & Beguin, 2002; Kim & Cohen, 1998; Kim & Cohen, 2002; Baker & Al-Karni, 1991; Lee & Ban, 2010; Hu, Rogers, & Vukmirovic, 2008). Only several comparisons studies have been conducted under the vertical scaling context (Karkee et al., 2003; Meng, 2007; 2003; Briggs & Weeks, 2009; Ito, Sykes, & Yao, 2008). In these studies, the TCC scaling method is the only transformation method that was used for the separate calibrations.

For studies conducted in the vertical scaling context, the different calibration methods produced somewhat different results and the conclusions are not consistent across studies. Karkee and colleagues (2003) concluded that the separate calibrations produced consistently better results than the concurrent and hybrid calibrations. Ito and colleagues (2008) found that concurrent calibrations and separate calibration yielded comparable results in reading tests and less comparable results in math tests. Meng (2007) and Briggs and Weeks (2009) had similar findings indicating that the hybrid

calibrations outperformed separate and concurrent calibrations. One possible reason for these discrepancies is that other vertical scaling factors are confounded with calibration methods. Therefore, additional comparisons studies are needed under different scaling conditions.

As described above, scale transformation is a critical step in both the separate calibration and hybrid calibration methods in vertical scaling. However, finding the most appropriate estimates of transformation parameters is challenging. First, there are many sets of linear transformation parameter estimates which can be used for scale transformation, but not all of them produce adequate linking results. Second, the scale transformation should be symmetric so that scale  $j$  can be transformed to scale  $k$  and scale  $k$  can also be transformed to scale  $j$ . This is different from non-symmetric simple regression techniques (Stocking & Lord, 1983). Third, a perfect scaling implies that for a given  $\theta$  the probabilities of a correct answer are equal over the range of the target metric scale for all common items in the two tests. Therefore, estimates of transformation parameters  $A$  and  $B$  must meet these criteria as closely as possible. In the following sections, the scale transformation methods for unidimensional IRT models are discussed.

### **2.3.3 Scale transformation methods for unidimensional model**

There are two general scale transformation approaches: moments methods and characteristic curves methods. The moments methods use the first two moments, mean and standard deviation, of the distributions of estimated item parameters to find transformation parameters  $A$  and  $B$ . The characteristic curves methods use more

information from the item characteristic curves (ICC, also referred to as IRF) or test characteristic curves (TCC, also referred to as TRF) to find appropriate values of  $A$  and  $B$ . In addition, the minimum  $\chi^2$  method (Divgi, 1985), the least squares method (Ogasawara, 2001), and ability transformations are also used by researchers to obtain scaling parameters.

#### 2.3.4 Moments methods

The moments methods use the first two moments of the distributions of estimated common item parameters to find transformation parameters  $A$  and  $B$ . More specifically, it uses either the mean or standard deviation, or both, of estimated  $a$ -parameters or  $b$ -parameters of the common items from two separate calibrations to estimate transformation parameters.

*Mean/mean method.* The mean/mean method (Loyd & Hoover, 1980) uses the means of the estimated common item parameters,  $a$  and  $b$ , to obtain the transformation parameters. The mean of estimated  $a$ -parameters for the common items is used to get the estimate of  $A$ . The mean of the estimated  $b$ -parameters of common items is used to obtain the estimate of  $B$ . Suppose there are two level scales,  $j$  and  $k$ , for common items, and the level  $k$  scale is transformed to level  $j$  scale. The parameters of  $A$  and  $B$  are estimated as:

$$A = \frac{\bar{a}_k}{\bar{a}_j}, \text{ and } B = \bar{b}_j - \frac{\bar{a}_k}{\bar{a}_j} \bar{b}_k. \quad (8)$$

*Mean/sigma method.* The mean/sigma method (Marco, 1977) only uses the information of  $b$ -parameter estimates from the common items to obtain estimates of  $A$  and  $B$ , rather than using the information from both  $a$ -parameter and  $b$ -parameter. The standard deviation of the estimated  $b$ -parameters for the common items is used to find the estimate of  $A$ . The estimate of  $B$  is obtained from both the mean and standard deviation of the  $b$ -parameters. Suppose there are two level scales,  $j$  and  $k$ , for common items, and the level  $k$  scale is transformed to level  $j$  scale. The estimates of  $A$  and  $B$  are:

$$A = \frac{s_j}{s_k}, \text{ and } B = \bar{b}_j - \frac{s_j}{s_k} \bar{b}_k \quad (9)$$

*Robust methods* One limitation of the mean/sigma method is that poorly estimated  $b$  parameters impact the distribution of  $b$ -parameter estimates and further distort the estimates of the scaling parameters  $A$  and  $B$ . Robust methods were developed to reduce the impact of poor or deviant item parameter estimates. The fundamental rule of robust methods is similar to that of the mean/sigma method. The main difference is that the robust methods use the weighted mean and sigma rather than using the geometric mean and sigma of  $b$ -parameter estimates.

In the robust method developed by Linn, Levine, Hastings, and Wardrop (1980), the weights are inversely proportional to the estimated standard error of the estimates of the  $b$ -parameter. In this procedure, the items with the same standard errors are treated in the same way, regardless of their status. That is, they ignore whether the estimates of the  $b$ -parameters are deviant or not. Bejar and Wingersky (1982) developed a similar weighted mean/sigma method, but they took not only the mean but also the median of

the  $b$ -parameter estimates into account for estimating  $A$  and  $B$ . The introduction of the median reduces the impact of the very deviant  $b$ -parameter estimates. In addition, they gave smaller weights to the outliers, regardless of the standard error of the estimates of the  $b$ -parameter. Ignoring of standard error of estimates may be a potential limitation of this robust method. Lord and Stocking (1983) developed an iterative robust mean/sigma method to overcome the potential problems of the weighted mean/sigma methods. The basic idea of this method is similar to other robust methods by introducing weights to item estimates.

### 2.3.5 Characteristic curve methods

The characteristic curve methods minimize a quadratic loss function that depends upon the metric of the test calibration (Baker & Al-Karni, 1991; Kolen & Brennan, 2004). It includes the item characteristic curve method (ICC or IRF) and the test characteristic curve (TCC or TRF) method.

*ICC method.* The ICC method (Haebara, 1980) minimizes the difference between the item characteristic curves of common items based on two calibrations. The difference in the ICC method is a squared difference between each pair of item characteristic curves for an item based on two calibrations for a given ability,  $\theta$ . The quadratic loss function is then accumulated over all  $v$  common items. Finally, the sum of the loss function is accumulated and averaged over all  $N$  examinees. It is expressed as (Baker & Al-Karni, 1991; Baker, 1996; Haebara, 1980; Kolen & Brennan, 2004):

$$H_{crit} = \frac{1}{N} \sum_i^N \sum_j^v [P_i(\theta|a, b, c) - P_i(\theta^*|a^*, b^*, c^*)]^2 \quad (10)$$



$P_i(\theta^*|a^*, b^*, c^*)$  and  $P_i(\theta|a, b, c)$  are the probabilities of correctly answering a common item  $i$  based on two calibrations. The estimates of the transformation parameters,  $A$  and  $B$ , are found by minimizing the above criterion.

*TCC method.* Similar to the ICC method, the TCC method (Stocking & Lord, 1983) also minimizes a quadratic loss function which is an average difference over all  $N$  examinees. However, the TCC method (Stocking & Lord, 1983) minimizes the difference between the test characteristic curves of common items based on two separate scales. Specifically, the difference in the TCC method is a squared difference between two test characteristic curves of  $v$  common items that are based on two scales for a given  $\theta$ . It is expressed as (Baker & Al-Karni, 1991; Kolen & Brennan, 2004; Baker, 1996):

$$SL_{crit} = \frac{1}{N} \sum_i^N [\sum_j^v P_i(\theta|a, b, c) - \sum_j^v P_i(\theta^*|a^*, b^*, c^*)]^2 \quad (11)$$

An iterative approach is then used to find the estimates of  $A$  and  $B$  based on minimizing this criterion.

### 2.3.6 Minimum $\chi^2$ method and least squares method

*Minimum  $\chi^2$  method.* Divgi (1985) developed a minimum  $\chi^2$  method which is based on the characteristic curves methods and robust methods. The basic idea of the minimum  $\chi^2$  method is also to minimize some measure of the difference of  $v$  common items. The  $\chi^2$  can be written as (Divgi, 1985; Kim & Cohen, 1995):

$$\chi^2 = \sum_j (a_j - a_j^*, b_j - b_j^*) (\Sigma_j + \Sigma_j^*)^{-1} (a_j - a_j^*, b_j - b_j^*)' \quad (12)$$

where  $\Sigma_j$  is the estimated 2x2 variance-covariance matrix of sampling errors for item  $j$  from the first calibration, and  $\Sigma_j^*$  is the transformed variance-covariance matrix from the second calibration;  $a_j$  and  $b_j$  are the estimates of discrimination and difficulty parameters from the first calibration, respectively;  $a_j^*$  and  $b_j^*$  are the transformed estimates of discrimination and difficulty parameters from the second calibration, respectively. The scaling parameters  $A$  and  $B$  are obtained by minimizing the above  $\chi^2$ . One primary advantage of this method is that equation (12) is directly related to parameter  $B$  (Divgi, 1985). Another advantage is that it takes the information not only from each item parameter, but also from the standard error of estimates. Therefore, it can be viewed as a variation of characteristic curves methods but with the standard error of estimates also taken into account. However, because the minimum  $\chi^2$  method is based on the assumption that  $\theta$  parameters are already known, the matrix is an underestimate (Ogasawara, 2001b).

*Least squares method.* The least squares method (Ogasawara, 2001b) was proposed to avoid an iterative computation, but the fundamental rule of the least squares method is similar to the minimum  $\chi^2$  method. The estimations of  $A$  and  $B$  are based on some least squares functions.

### 2.3.6.1 Proficiency transformation method

The proficiency transformation method is a newly proposed scale transformation method. It is also called a common population linking method (Lee, Song, & Kim, 2004;

Lee & Ban, 2010). It uses the estimated ability distributions to find scaling parameters rather than using the estimated item parameters. For a given group of examinees, different estimates of the same group of examinees can be obtained based on different calibration scales. According to item response theory, the linear parameters for transforming the abilities on one scale to abilities on the other can be estimated by setting the standardized estimates (or z-scores) of ability as equal (Marco, 1979). Given an examinee, the ability estimates on two different scales,  $\theta_1$  and  $\theta_2$ , have the following relationship:  $\theta_1 = \frac{s_1}{s_2}\theta_2 + \bar{X}_1 - \frac{s_1}{s_2}\bar{X}_2$ , where  $\sigma_1$ ,  $\bar{X}_1$ ,  $\sigma_2$ , and  $\bar{X}_2$  are the standard deviation and mean for two sets of ability distributions. Then, the transformation parameter  $A$  and  $B$  can be expressed as:  $A = s_1(\theta)/s_2(\theta)$  and  $B = \bar{X}_1 - A * \bar{X}_2$ . The proficiency transformation method is proposed as an alternative to the scaling transformations based on item parameters. However, the ability parameter estimates could be unstable if there are only a few common items or if some common items are sensitive to different test situations. There is limited research on this scaling transformation method.

### **2.3.6.2 Comparison studies of scale transformation methods**

As described above, a variety of scaling transformation methods can be applied. The mean/mean method and mean/sigma method only use summary statistics of item parameters. While this statistical simplicity is an attractive feature of the moments method, it is also a potential limitation since not all available information from the item parameters is used simultaneously. Ignoring some information may produce inadequate results. For the robust moments methods, including weights reduces the impact of very deviant  $b$ -parameter estimates. Including the standard error of estimates further improves

the estimation of  $A$  and  $B$ . However, the robust methods involve complex computations and ignore the information from the  $a$ -parameter.

The salient feature of characteristics curves methods is that all available information from item parameters ( $a$ ,  $b$ , and  $c$  parameters) is used. Theoretically, the more information that is included, the better the estimates. However, the characteristic curve methods require complex iterative multivariate searching to obtain the estimates. In addition, the characteristic curve methods do not consider the standard error of estimates, which might lead to problems in estimating  $A$  and  $B$  (Baker & Al-Karni, 1991).

The Minimum  $\chi^2$  method is simpler than the characteristic curve methods and makes more complete use of available information that includes item parameter estimates and the standard error of estimates (Divgi, 1985; Kim & Cohen, 1995), but it still needs an iterative computation. More importantly, the variance-covariance matrix which is used for estimation is an underestimate. This underestimation may lead to biased scaling results. The least squares methods are similar to the minimum  $\chi^2$  method. These two methods are less attractive than moment methods and characteristic curves methods due to their statistical complications.

Stocking and Lord (1983) compared the TCC to the robust mean/sigma method by using scatter plots. The results showed that the robust mean/sigma method never provided a better fit to the estimated  $a$ -parameters and  $b$ -parameters as compared to the TCC method. They concluded that the TCC method was logically superior to the robust mean/sigma method because the TCC uses more available information from item parameter estimates than the robust mean/sigma method. However, this superiority does not always hold when using long chains of transformations. This is most likely due to

item sampling fluctuations. Haebara (1980) reported that the ICC method was more accurate in recovering true values of linking coefficients for the 3P logistic model than mean/sigma model.

In a study by Baker and Al-Karni (1991), the mean/mean method and the TCC method were compared under three types of testing situations: IRT parameter recovery, horizontal equating, and vertical scaling. The results showed that the two scaling methods generally yielded similar transformation coefficients in all testing situations. There was also little difference between these two methods when using an actual data set. In general, the mean/mean method produces acceptable scaling results. The TCC method behaves a little bit better than the mean/mean method and appears to be less sensitive to atypical test characteristics. Therefore, the TCC may be preferred when the test data are suspected to be atypical.

Hanson and Béguin (2002) conducted a more comprehensive comparison among the scaling methods. In a simulation study, the mean/mean, mean/sigma, ICC method, and TCC method were compared using a common-item equating design. They found that the ICC and TCC methods had lower mean squared error (MSE) of estimated item characteristics than the mean/mean and mean/sigma methods. The lower MSE of the characteristics curve methods is primarily due to the lower variance of the item parameter estimates. The MSEs for the ICC and TCC methods are similar to each other, and neither method had consistently lower MSEs than the other. When the two test forms are non-equivalent, the bias of the mean/mean method is higher than the mean/sigma method. Otherwise, the two methods produced similar results. The results suggested that the characteristic curve methods for item parameter scaling should be preferred over the moment methods, which is consistent with previous studies.

Kim and Kolen (2006) compared the concurrent calibration method to the four scaling methods, mean/mean, mean/sigma, ICC, and TCC, with the focus on the robustness to mixed-format effects. Format effects occur when the examinees process items differently due to the format of the items. For example, the examinees' performance may be not consistent on multiple-choice items and constructed-response items, even though these two types of items may be constructed to measure the same proficiency. When format effects occur, it will lead to the presence of multidimensionality at the level of total test scores (Kim & Kolen, 2006).

In their study, three factors were manipulated: (a) three levels of format effects, (b) two types of mixed-format tests, (c) three levels of nonequivalence between two examinee groups to be linked. The levels of format effects were manipulated by the correlation between the proficiency on multiple-choice items and the proficiency on constructed-response items. The correlations were set at 1, 0.8, and .5. The two types of mixed-format tests were specified as a wide range test and a narrow range test. The definition of the two types of mixed-format tests was based on the test information at proficiency levels. The target information peaked at .5 and 1.0 for the wide range test and narrow range test, respectively. The levels of nonequivalence between two examinee groups were manipulated by adding different values to the latent ability.

Their results indicated that the concurrent calibration outperformed separate calibration with four linking methods; the characteristics curves methods produced more consistent and stable results than moments method. This finding should not be generalized to a vertical scaling since only two nonequivalent examinee groups were investigated.

Lee and Ban (2010) used a simulation study to compare different linking and calibration procedures in a random groups equating design. The equating design in this study had three examinee groups, group 1, group 2, and group 2'. Groups 1 and 2 took the same test form and group 2' took a different test form. Group 2 and group 2' were equivalent groups. In their study, the sampling conditions, the sample size, and the number of total items were varied. The sampling conditions were manipulated by setting different means of ability of three examinee groups. For example, if group 1 and group 2 were not equivalent, the means of ability for group 1, group 2, and group 2' can be set at 0, 0.5, and 0.5, respectively. This is one sampling condition. Three combinations of sampling conditions were investigated: M0/0/0, M0/0.5/0.5, M0/1/1. The two levels of sample size were 3000 and 500. The number of total items varied at two levels: 75 and 25.

The study results showed that the separate calibration procedures performed better than the concurrent calibration and proficiency procedures. ICC method produced lower linking error than the TCC method. However, when the three samples were from the same population, concurrent calibration outperformed the separate calibrations and proficiency transformations. One limitation of this study was that scaling transformations were not needed if the two groups were truly equivalent.

Karkee and his colleagues (2003) examined the concurrent calibration, the separate calibration with TCC transformations, and pair-wise calibration methods under the vertical scaling context using operational test data. The operational test data were from the Colorado Student Assessment Program (CASP, 2002) math test results. The 2002 CASP math scores were placed on a common scale spanning grades 5 through 10 by using a common-items nonequivalent groups design. The item numbers for each

grade test varied from 60 to 70 with about 20 items in common between adjacent grades. The study results indicated that separate calibration produced consistently better results than the concurrent calibration method and pair-wise calibration method.

In general, concurrent calibration and separate calibration are the two most often used calibration methods. Mean/mean, mean/sigma, ICC, and TCC are the common scale transformation methods when using separate calibration. One advantage of the mean/mean and mean/sigma method is their statistical simplicity. The difference between mean/mean method and mean/sigma method is small. The comparison studies suggest that characteristics curves methods may produce slightly better results than the moments methods. The difference between the ICC and TCC method is relatively small even though the TCC is more often used by researchers.

### **2.3.7 Comments about choosing scale transformation methods**

As mentioned earlier, there are no theoretical criteria for choosing a particular calibration and scaling method. Usually, the selection is based on both previous comparison studies and the features of the methods themselves. For example, the moments method, such as the mean/sigma method, may be preferred because of its statistical simplicity. However, if the test is suspected to have very atypical item parameter estimates, the use of the mean/sigma method may be questionable since it is highly sensitive to item parameter outliers.

It is important to note that results from empirical comparison studies are usually conditional and may not generalize to other conditions. That is, in the comparison studies, only limited linking or equating factors are investigated. Different conditions



likely yield different conclusions about the behavior of scaling methods. Thus, the generalizability of comparison results is limited.

In the vertical scaling design, choosing and implementing scale transformations requires more attention. First, for the widely used IRT-based vertical scaling, unidimensionality is a fundamental assumption. However, in practice, this assumption is very likely to be violated when scaling is conducted across a wide range of ability levels.

Second, the common-items design itself has limitations. In a common-item non-equivalent groups design, the adjacent groups take the same set of common items. There is typically a mismatch for at least one group of examinees and the difficulty of the common items. When there is a mismatch, the error in the item parameter estimates is larger than when the test matches examinee ability. Usually, the scaling methods are more accurate when there is less error in the item parameter estimates. If there is only small to moderate error in item parameter estimates, the TCC method is relatively robust (Kaskowitz & De Ayala, 2001). Therefore, the difficulty of the common-items should match the examinees in the two adjacent grades as close as possible.

In addition, the scaling results could be biased if there are very atypical common items or some items with differential item functioning (DIF). Therefore, it may be reasonable to exclude very atypical items and items with DIF when estimating parameters  $A$  and  $B$ . However, DIF may be a source of interested multidimensionality.

Overall, there are no theoretical criteria for choosing a particular scaling transformation method in practice. The selection depends on previous comparison studies, the characteristics of scaling methods, and the features of the real data. Hanson and Béguin (2002) suggested that it would be beneficial to apply multiple linking procedures and compare the scaling results. This will allow for a better understanding of

the various issues and aspects of the scaling situation and will inform the choice of the scaling method.

## **2.4 ABILITY ESTIMATION**

In IRT-based vertical scaling, once the item parameters are calibrated on a common scale, the examinees' latent abilities are then estimated on a common metric. The IRT-based ability estimates (also referred to as proficiency estimates) identify an examinee's location ( $\theta$ ) on a latent-trait continuum by using an examinee's item response pattern in conjunction with the estimated item parameters (Embretson & Reise, 2000). IRT-based proficiency estimation methods include maximum likelihood estimation (MLE) and two Bayesian methods, maximum a posterior (MAP) and expected a posterior (EAP) (Kolen & Brennan, 2004; Tong & Kolen, 2007; Briggs & Weeks, 2009; Briggs & Weeks, 2008).

### **2.4.1 Maximum likelihood estimation (MLE)**

MLE is a search process that finds the value of  $\theta$  that maximizes the likelihood of an examinee's item response pattern. More specifically, an MLE estimate of the latent trait is determined by summing the likelihood of each observed item response conditional on the value of  $\theta$ . Then, the maximum or the mode of the likelihood function is found by numerical methods such as the iterative Newton-Raphson procedure or EM procedure. Given local independence, the conditional likelihood function of an examinee's item response pattern (also referred as the joint probability of the responses to a set of  $I$  items)

is the product of item response curves across all administered items (Embretson & Reise, 2000):

$$L(u_1, u_2, \dots, u_i | \theta) = \prod_{i=1}^I P_i(\theta)^{u_i} P Q_i(\theta)^{1-u_i} \quad (13)$$

where  $u_i$  is the response to the  $i$ th item;  $\theta$  is the ability level on the underlying latent-trait continuum;  $P_i(\theta)^{u_i}$  is the probability of correct response of item  $i$  conditional on the ability  $\theta$ ;  $Q_i(\theta)^{1-u_i}$  is the probability of incorrect response of item  $i$  conditional on the ability  $\theta$ .

#### 2.4.2 Maximum a posterior (MAP) and Expected a posterior (EAP)

Both MAP and EAP are Bayesian estimates which incorporate prior information about latent traits to examinees' item responses. Incorporating information about a prior distribution allows for more efficient estimates and protects against outliers or influential data points that may have undue influence on ability estimates.

The Bayesian estimates are derived from the posterior distributions which makes the use of prior information about the ability levels in conjunction with the observed log-likelihood function (Embretson & Reise, 2000; Stone, 2007):

$$P(\theta|u) = L(u|\theta)g(\theta)/[\sum_{j=1}^J L(u|\theta_j) * g(\theta)] \quad (14)$$

where  $P(\theta|u)$  is the posterior distribution of latent ability  $\theta$  conditional on response pattern  $u$ ;  $L(u|\theta)$  is the likelihood function for the responses pattern  $u$ ;  $g(\theta)$  is the assumed prior distribution of  $\theta$  which is usually assumed to be a normal distribution with mean 0 and variance 1;  $[\sum_{j=1}^J P(u|\theta_j) * g(\theta)]$  is the marginal probability distribution of response pattern  $u$  across  $\theta$ .

EAP estimates use the mean of the posterior proficiency for a given response pattern while MAP estimates use the mode of the posterior distribution. MAP estimates can be obtained in the same way as MLE estimates. EAP estimates are obtained by an integration algorithm for a continuous distribution. In practice, for statistical simplicity, it uses the summation of a set of  $k$  discrete values for  $\theta$  instead of using integration. The EAP estimates can be expressed as:

$$\text{EAP}(\theta) = \frac{\sum X_k L(u|X_k) g(X_k)}{\sum L(u|X_k) g(X_k)} \quad (15)$$

where  $X_k$  is the  $k$ th discrete value for  $\theta$ ;  $L(u|X_k)$  is the log likelihood function of response pattern  $u$  given the  $k$ th discrete value for  $\theta$ ;  $g(X_k)$  is relative probability of  $\theta$  from the prior distribution.

### 2.4.3 Comparisons of ability estimation methods

Theoretically, MLE estimates are unbiased estimates for a test with reasonable length. They are also efficient estimates and their errors are normally distributed. However, the MLE method provides infinite estimates for examinees with extreme abilities. In other

words, no ability estimate exists for students who have item response patterns with all 0s or all 1s. In practice, the estimates for all 0s or all 1s need to be arbitrarily assigned such as -5 and +5, respectively. However, the arbitrary assignment likely distorts the accuracy of the estimated ability distribution.

One salient feature of EAP and MAP is that they produce finite estimates for all examinees. In addition, both EAP and MAP yield smaller standard errors than MLE. In contrast to MAP, the solution to EAP estimates is a noniterative procedure and thus can be easily computed. However, the ability estimates of Bayesian methods depend not only on examinee's test performance but also on the nature of the entire group in which she or he belongs. Therefore, the estimates derived from Bayesian methods are theoretically biased estimates. The estimates are likely regressed toward the mean unless the sample size is large (Embretson & Reise, 2000). In general, one critical tradeoff between MLE and Bayesian methods is one of efficiency versus bias.

The choice of proficiency estimation methods is relatively new in vertical scaling research. A few of studies have been conducted to compare how the Bayesian and MLE methods behave in the vertical scaling context (Tong & Kolen, 2007; Briggs & Weeks, 2009). The results of these studies indicate that the ability estimation method is an important factor related to vertical scales and growth patterns.

Tong and Kolen (2007) compared several IRT proficiency estimation methods such as MAP, EAP, MLE, and quadrature distribution (QD) for both real data and simulated data under the vertical scaling context. QD estimates the entire "true" proficiency distribution, and it does not provide estimates for individual proficiency. The test data of the Iowa Test Battery (ITBS) in 1992 from grades 3 to 8 were used for the real data study. Four test content areas were investigated: vocabulary, math, language,

and reading. The ITBS scales were constructed by using a scaling test design. The results of the real data study indicated that the MAP, EAP, and QD ability estimations methods produced similar results. The MLE estimates were slightly different from the other three ability estimates.

For the simulation study, the 3P logistic model was used as the IRT model. The data were generated under both a scaling test design and a common-items nonequivalent groups design. Grade 3 was used as the base grade. Three levels of sample size were manipulated: 500, 2,000, and 8,000. A separate calibration was used to put item parameter estimates on a common scale for the IRT-based scaling. In the scaling test design, proficiency transformation was used as the linking method. In the common-items design, the TCC method was used to transform item parameters. The evaluation criteria included estimated means, within-grade variability, and effect size (*e.g.*, year-to-year growth).

The results were complex under different conditions with different evaluation criteria. For estimated means, the MAP, EAP, and MLE yielded similar results. For within grade variability, the MLE method likely overestimated the within grade variances; the MAP method likely underestimated the within grade variances; and the within grade variance of EAP estimates was close to the “true” variance. For effect size, the EAP and the MAP methods tended to exhibit the similar growth across the grades. In general, EAP and MAP procedures tended to produce more accurate estimates than the MLE procedure under the simulated data in the vertical scaling contexts.

Briggs and Weeks (2009) also compared the EAP and MLE methods using real data under the vertical scaling context. The data were from the Colorado Student Assessment Program (CSAP) reading test from 2003 to 2006. The CSAP vertical scale

was based on a 3P logistic model and a common item nonequivalent group design. In the study, they created and compared different vertical scales by manipulating three factors: (1) the IRT models used to estimate item parameters, 1P model versus 3P model, (2) the calibration and linking methods, separate calibration with TCC linking method (or SL method) versus hybrid calibration method, (3) ability estimation methods, the EAP method versus MLE method. The evaluation criteria were means, standard deviations, and effect sizes.

The study results indicated that the 3P model with separate calibration yielded the most growth. The 1P model yielded the least growth regardless of the ability estimation methods. The EAP method yielded larger growth for students than the MLE method. One limitation of this study is that the 1P model may not fit the data as well as the 3P model. In addition, all of the vertical scales showed patterns of decelerating growth across grade levels. They also concluded that the empirical growth patterns appear to depend on different vertical scaling methods.

In general, these studies showed that Bayesian methods perform better than MLE method. One possible reason is that the latent ability distributions cross a really wide range. The ability estimates by using MLE become very unstable in the tails of the ability distribution. Therefore, one of the two Bayesian methods, the MAP procedure, is selected as the ability estimation method in this study.

## **2.5 FACTORS RELATED TO VERTICAL SCALING**

### **2.5.1 Number of grade levels and choice of a base year**

When designing a vertical scale, the number of grade levels and choice of a base year are two important issues that need to be taken into account (Chin, Kim, Nering, 2006; Smith, *et. al.*, 2008). When the number of the grades increases, the scaling errors increase in the separate calibrations since more transformations are needed. Also the bias accumulates when using a long chain process. For the concurrent calibration, the unidimensionality assumption is more likely to be violated when the number of grades increases. In other words, regardless of the calibration method, the errors are larger when more grades are to be scaled. However, in practice, scores from a relatively large number of grades need to be scaled. In general, six grades is a typical and reasonable number for vertical scaling (Karkee, *et. al.*, 2003; Tong & Kolen, 2007; Briggs & Weeks, 2009).

In IRT-based vertical scaling, a base year needs to be defined before conducting the calibrations. The scale of the base year is used to put the item parameter estimates of other grades on that scale. Some studies used a middle grade as the base grade (Karkee *et. al.*, 2003; Briggs & Weeks, 2009; Ito, *et. al.*, 2003), while other studies used a beginning year as the base grade (Tong & Kolen, 2007, Yin & Stone, 2009). The last year also could be the base year. There are few comparison studies examining the impact of using the beginning year or the middle year as the base grade on vertical scaling results (Kim, Lee, Kim, & Kelley, 2009). According to research (Briggs & Weeks, 2009; Ito, *et. al.*, 2003; Tong & Kolen, 2007; Yin & Stone, 2009) the bias and standard errors



of ability estimates and scaling results are smaller for the base year than the other years and the errors get larger as the grades depart from the base year.

Kim, Lee, Kim, and Kelley (2009) used different grades as the base grade to construct vertical scales from grades 4 to 8 using a Rasch model. In the simulation study, the number of common items (approximately 50% and 25% of a full-length test) and sample size (250, 500, and 1000) were manipulated. The separate calibration with mean/mean, mean/sigma, Stocking-Lord transformations, and fixed parameter calibration were also compared. They found that the mean squared errors (MSE) of ability estimates were smaller when the sample size and number of common items were larger.

Overall, they found the MSE of ability estimates for the base grade was much smaller regardless of which grade was selected as the base grade. The MSE of ability estimates was the smallest when the middle grade, grade 6, was used as the base grade. Also, the results showed that the MSE increased as the base grade moved away from the grade of interest, indicating that, as the number of linkings increased, the errors of ability estimates accumulated. Thus, the middle year grade may be preferred as the base grade in a vertical scaling design since it reduces the errors by breaking a long chain transformation into two parts, from the middle to the beginning year and from the middle to the last year. Therefore, in this study, six grades (grades 3 to 8) are used to conduct a vertical scale and a middle year, grade 5, is selected as the base grade.

### **2.5.2 Sample size**

Identifying and acquiring a large enough representative sample of examinees is crucial to vertical scaling. The sample size directly affects the item parameter estimates. Theoretically, the larger the sample size, the more accurate the scaling results. If the sample size is too small, the item parameter estimates are unstable. However, in practice, a larger sample size is difficult to obtain in vertical scaling. Liu and Walker (2007) suggested that the identification of appropriate sample size is mainly based on three criteria: adequate representation to ensure statistical precision, adequate representation of subgroups of the population, and economic considerations.

According to Swaminathan and Gifford (1983) (referred to in Fitzpatrick & Yen, 2001), 1,000 cases are needed to produce adequate parameter estimates when 3P the logistic model is used. When the Rasch or 2P logistic IRT models are used, the requirements for the sample size are smaller. Liu and Walker (2007) show that the minimum sample size is about 975 to produce stable results for the 3P logistic model. In summary, for the 3P logistic IRT model, about 1000 examinees in each grade can produce adequate estimations.

### **2.5.3 Test length and number of common items**

The quality of item parameter estimates is affected by the test length in combination with the sample size (Fitzpatrick & Yen, 2001). Theoretically, the longer the test, the more accurate the results. In practice, the length of multiple-choice tests should be chosen based on the test blueprint and on the characteristics of the students so that the

students can finish the test in an appropriate time period. For example, the test should be shorter for lower grades and longer for higher grades. Typically, the number of dichotomous items in a test range from 45 to 60 (Fitzpatrick & Yen, 2001).

In a common-item design, the number of common items is also a significant factor that affects scaling results. Typically, the larger the common item set, the smaller the mean squared error (Kolen & Brennan, 2004). However, if the total test length is fixed, more common items imply less grade level items. This might affect the reliability of the test scores. There is no consensus on the number of common items required to provide adequate linking. In practice, for horizontal equating, the number of common items of a test usually is set at no less than 20% of the test length of 40 items or more (Kolen & Brennan, 2004; Kim & Kolen, 2006). For example, if there are 50 items, the number of common items should be at least 10.

## **2.6     DIMENSIONALITY ISSUES**

### **2.6.1   Unidimensionality assumption**

One fundamental assumption for the logistic IRT model is local independence. It assumes that for examinees with the same ability, the probability of getting a given item correct is independent of the performance on any other items in the same test. However, if the latent construct being measured is actually multidimensional, this assumption may be violated. This violation will further bias item and ability parameter estimates, and the standard errors associated with ability estimates will be very small (Briggs, 2008).

Therefore, it leads to another important assumption, unidimensionality or essential unidimensionality.

The unidimensionality or essential unidimensionality of a test is critical in IRT-based vertical scaling. It assumes that item responses are based on the same skill or same composite of multiple skills (Walker, Azen, & Schmitt, 2006; Ackerman, 1994). The unidimensionality assumption mainly includes two aspects: homogeneity of content and construct invariance (Reckase & Martineau, 2004). Homogeneity of content means the tests should measure a similar content domain across grades. Construct invariance means the weights or number of items in each sub-construct in the broad construct should be parallel across the grade level tests. As an example, the weights of different specific topics such as algebra and geometry in a math achievement test should be similar across multiple grades. These assumptions are critical in vertical scaling since growth cannot be determined if the content and the weights of subtests are different. However, in practice the weights of subtests differ across grades.

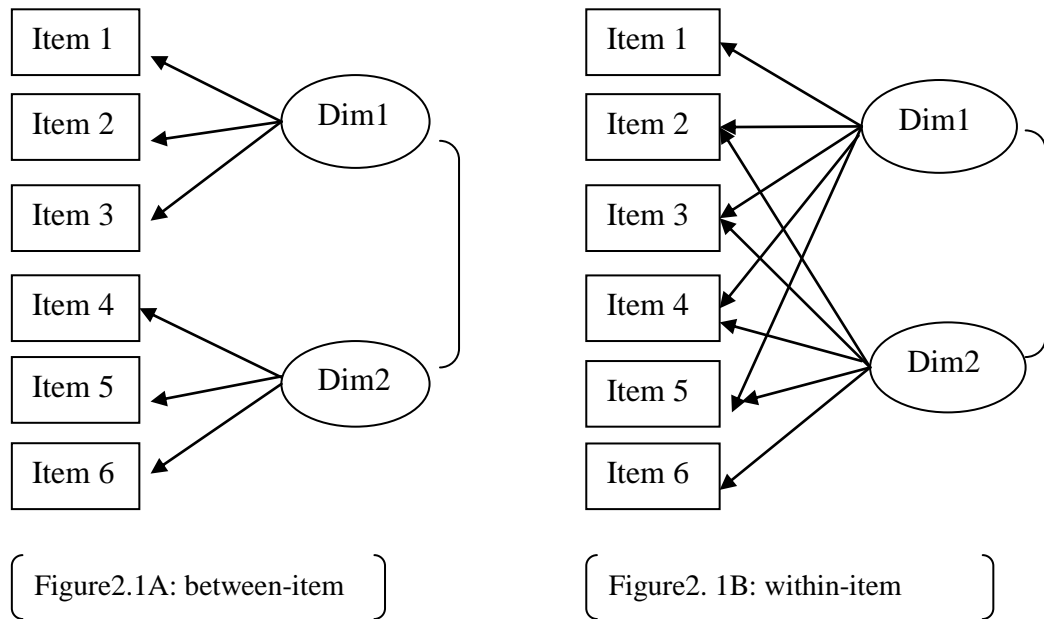
Schmidt, Houang, and Mcknight (2005) used the data from the Third International Math and Science Study (TIMSS) and General Topic Trace Mapping method (GTTM) to analyze cross-level tests on the assumption of homogeneity of content. The GTTM is a new measurement technology to track the test content topic across grades. The GTTM provides broad information on all of the topics of the content aspect of TIMISS. The results indicated that the homogeneity of content was likely violated in cross-level test.

Construct invariance is another condition that must be met for appropriate use of vertically scaled cross-level tests. Depending on the nature of measured construct, some specific and prominent skills in the general content domain are usually associated with a

particular grade level. In test development, the different prominent skills typically have different construct weights in cross-level tests (Lockwood, *et al.*, 2007). Martineau (2006) used a mathematical approach to investigate if there is violation of measurement invariance in a vertically scaled developmental score. The results showed that there was a shifting in the construct across wide grade ranges. That is, there were violations of measuring similar content and the same weight of each main topic across grades. The violation of unidimensionality is considered as a possible reason for scale shrinkage (Camilli *et al.*, 1993) and a lack of equating invariance (Skaggs & Lissitz, 1988). When there is a violation of unidimensionality of a test, the test usually is viewed as exhibiting a multidimensional data structure.

### **2.6.2 Multidimensional data structures**

There are two main types of multidimensional data structures (Figure 2.1). One is the between-item multidimensional data (Figure 2.1A) and the other is the within-item multidimensional data (Figure 2.1B) (Briggs, 2008; Wilson, 2009):



**Figure 2.1** Multidimensional data structure

Between-item multidimensionality exists when each item measures one dimension while the whole test measures two or more dimensions. Usually these abilities are moderately to highly correlated with each other. This type of data structure is also called simple structure. For example, in a general math skill test as indicated in Figure 1A, items 1, 2, and 3 measure an ability related to a specific topic such as algebra. Items 4, 5, and 6 measure an ability related to another topic such as geometry. The test thus has a between-items multidimensional data structure.

Within-item dimensionality exists when an item measures several latent abilities all together and two or more dimensions contribute significantly to students' scores. These dimensions are also likely correlated with each other. The correlations can vary from low to high. As indicated in Figure 1B, items 1 and 6 measure two different abilities, but items 2, 3, 4, and 5 measure two abilities simultaneously. This type of data

structure is more complex than the between-item multidimensional data structure. Therefore, it is sometimes called non-simple structure or complex structure.

As mentioned earlier, the assumption of unidimensionality is not truly met for many tests when the tests are designed across grade levels (Reckase & Martineau, 2004; Yao & Mao, 2004; Lockwood, *et.al*; Martineau, 2006). Instead, the tests are likely to exhibit a multidimensional data structure across grades.

### **2.6.3 Effects of multidimensionality on IRT-based vertical scaling**

Considerable attention has been given to studying the effects of multidimensionality on IRT-based equating results. De Champlain (1996) attempted to assess whether differences in the dimensional structure of two test forms across three ethnic subgroups had any impact on IRT-true equating functions. The equating results indicated that the difference between subgroups were negligible throughout the entire raw-score scale except for at the low end of the score scale.

Béguin, Hanson, and Glas (2000) investigated the effects of multidimensionality on separate and concurrent estimation under IRT equating. They used a simulation study to compare the relative performance of unidimensional estimation methods (concurrent and separate) on multidimensional data. Data were simulated under both equivalent groups and non-equivalent group designs by using a two-dimensional 3P normal ogive model. To explore the effects of multidimensionality, they used a 3P logistic unidimensional IRT model. The results showed that the unidimensional models yielded reasonable estimated score distributions when applied to multidimensional data under the equivalent group design. Under the non-equivalent group design, the deviation from

true score distributions was large. The effects of multidimensionality increased with larger covariance between two dimensions and this increase became larger for concurrent estimations. One limitation of this simulated study is that it used different types of IRT models for simulating data and estimating data. The differences between the logistic model and normal ogive model could lead to some variations between true and estimated score distributions.

There is limited research on the effects of multidimensionality on IRT-based vertical scaling. One possible reason is that the factors related to vertical scaling are much more complicated than equating. Another possible reason is that the multidimensionality structure in vertical scaling tests could be much more complex than the data structure in equating.

Smith, Finkelman, Nering, and Kim (2008) used a simulation study to compare five linking methods with unidimensional and multidimensional data under a vertical scaling design. The simulation study was designed to use a vertical scale based on an equivalent groups design for grade 3 through 8 with grade 5 as the base grade. In the simulation study, the number of total items for each grade was 60 and 25% of the items were used as linking items. That is, each test contained 45 grade level items and 15 linking items. Each grade (except grades 3 and 8) had three test forms because of the different linking items. For example, fifth grade students took one of the three test forms. For one form the 15 linking items were grade 4 items; for the second form the linking items were grade 5 items; and for the third the linking items were grade 6 items. The linking methods for vertical scaling were mean/mean method, mean/sigma method, the SL method, the Haebara method, and fixed common item parameter method. Both unidimensional data and multidimensional data were simulated.



To simulate multidimensional data, they used a simple MIRT structure and assumed that students had a true ability for each grade level items and a true ability for linking items and these two abilities were correlated. In the study, both the adjacent lower grade level items and adjacent higher grade level items were used as linking items to link adjacent grades. They simulated both unidimensional and multidimensional conditions. For the unidimensional condition, the linking results showed that the performance of different linking methods were very similar. For the multidimensional condition, the five linking and scaling methods also produced similar scaling results. However, some differences existed between the scaling results with unidimensional and multidimensional data.

In the study, when simulating the multidimensional data, the researchers assumed an “on grade” true ability for grade level items and a different mean ability for the linking items and they were correlated. The mean ability for common items were either higher for the common items from the adjacent lower grade or lower for the common items from the adjacent higher grade. For example, for grade 5 students, their on-grade ability level was set as:  $\bar{\theta}_5 = 0$ ; the ability for adjacent grade 6 items was set as:  $\theta_6 = +0.5$ ; and the ability for adjacent grade 4 items was set as:  $\bar{\theta}_4 = -0.5$ . The resulting data were then viewed as multidimensional data.

One limitation of the study was that the simulated multidimensional data structure only involved simple structure and the multidimensional data was questionable. In addition, the researchers did not mention whether they controlled the correlation between two latent abilities.

## 2.7 SUMMARY

Vertical scaling is very complex since it is usually developed across a wide range of grades and involves various processes and scaling methods. Many factors influence scaling results. These factors include the vertical scaling test design, the grades included in the scaling, the selection of base grade, test length and proportion of common items, sample size, calibration and scaling methods, ability estimation procedures, and dimensionality of data. These factors are likely confounded with each other.

Several studies have been performed to investigate different ability estimation methods, calibration methods, and growth patterns under vertical scaling when the assumptions generally hold (Tong & Kolen, 2007; Briggs & Weeks, 2009). In the Tong and Kolen study (2007), the focus was on the comparison of various ability estimation methods under different test conditions. Briggs and Weeks (2009) used a real test to investigate the impact of vertical scaling decisions, such as selecting an ability estimation method and a calibration method, on growth patterns. In general, research shows that vertical scaling is design-dependent, group-dependent, and method-dependent (Kolen & Brennan, 2004; Wang & Harris, 2009). Research suggests that it would be beneficial to apply multiple linking procedures and compare the scaling results in multiple situations.

There is limited research that has investigated the various calibration methods and scale transformation methods under the vertical scaling context. Furthermore, there is a paucity of research that has investigated the robustness of various vertical scaling methods to the violation of IRT assumptions, especially to the inevitable violation of the unidimensionality assumption. The investigation of the robustness of various vertical

scaling methods would further help understand various issues and aspects of the scaling situation and would inform the choice of the most appropriate scaling method in practice.

### **3.0 METHODS**

Vertical scaling is a complex statistical procedure whose outcome is design and method dependent. IRT-based vertical scaling is even more complex since it involves additional model assumptions. The purpose of this dissertation is to use Monte Carlo simulations to investigate the robustness of various unidimensional calibration and scaling procedures to the violation of unidimensional assumptions while developing vertical scales. Monte Carlo studies (also referred to as simulation studies) are an important means of evaluating new methods within psychometric research, particularly with respect to item response theory (IRT) models (Harwell, Stone, Hsu, & Kirisci, 1996).

In this chapter, the following methodologies for the simulation study are introduced: 1) Design of the simulation study; 2) Evaluation criterion; 3) Generation of data and program development.

#### **3.1 DESIGN OF THE SIMULATION STUDY**

##### **3.1.1 Vertical scaling data collection design**

In this simulation study, the common-items non-equivalent groups design was used since it is the easiest and most widely used vertical scaling design. Vertical scales are

constructed spanning from grade 3 to grade 8. Grades 3 through 8 were selected because these grades are involved in many assessment programs with vertical scales (Kolen & Brennan, 2004; Reckase & Martineau, 2004; Smith, *et. al.*, 2008 Tong & Kolen, 2007).

As mentioned earlier, common items are the items overlapping between any pair of adjacent grades and are used as anchors for the scaling. According to the different types of common item blocks, three specific linking designs are commonly used: below-grade/on-grade design, on-grade/above-grade design, and below-grade/on-grade/above-grade design, respectively (Wang, Jiao, Young, & Jin, 2005). In the below-grade/on-grade design, the common items are from the lower adjacent grade; in the on-grade/above-grade design, common items are from the higher adjacent grade; and in the below-grade/on-grade/above-grade, common items are from both the lower adjacent grade and the higher adjacent grade. Among them, the below-grade/on-grade design may be more preferable than the other two since it minimizes random guessing. If students take items in a higher grade level, they are more likely to randomly guess. Therefore, the below-grade/on-grade design was the specified common-items linking design in this study and is displayed in Figure 3.1.

Grades	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8
Grade 3	G3_on	G3_4c									
Grade 4		G3_4c	G4_on	G4_5c							
Grade 5				G4_5c	G5_on	G5_6c					
Grade 6						G5_6c	G6_on	G6_7c			
Grade 7								G6_7c	G7_on	G7_8c	
Grade 8											G7_8c

**Figure 3.1** Common-items non-equivalent groups design.

In Figure 3.1, item block G3\_on consists of the items appropriate for grade 3 students. Some of items from item block G3\_on are selected to make up the common items block, G3\_c. The items in the common block are then used to link the two adjacent grades 3 and 4. Therefore, grade 4 students take on grade level items, item block G4\_on, and some lower level items, *i.e.*, item block G3\_c. Grade 5 students take both the grade 5 on level item block and a common item block, G4\_c, which are from the grade 4 on level test. Similarly, grade 6, 7, and 8 students take both on grade level items and some common items from the lower adjacent grade.

In this study design, grade 5 was set as the base grade level with ability distribution of  $N(0,1)$ . According to empirical vertical scaling research (Karkee *et. al.*,

2003; Briggs & Weeks, 2009; Ito, *et. al.*, 2003; Tong & Kolen, 2007), the errors of item parameter recovery and scaling results are smaller at the base year and the errors are larger when the grades are far away from the base year. This implies that the middle grade may be preferred as a base in a vertical scaling design since it reduces the errors by breaking a long chain transformation into two parts (from the middle to the beginning year and from the middle to the last year). In this study, a total of six grades were used to conduct a vertical scale and the middle grade, grade 5, was selected as the base grade (Ito, *et. al.*, 2003; Smith, *et.al.*, 2008).

### 3.1.2 IRT model

The purpose of this study was to investigate the robustness of various unidimensional calibration and scaling procedures to the violation of the unidimensional assumptions while developing vertical scales. The Multidimensional IRT model was first utilized to simulate non-unidimensional data for common items, but only the unidimensional IRT model was used for item and ability parameter estimations.

The two-dimensional 3P logistic IRT model was used for item response generation only for common items. The two-dimensional 3P logistic IRT model is a compensatory item response model, whose form usually is expressed as (Reckase, 1985):

$$P_i(\theta|a, d, c) = c + (1 - c) \frac{e^{D(a_1\theta_1 + a_2\theta_2 + d)}}{1 + e^{D(a_1\theta_1 + a_2\theta_2 + d)}} \quad (16)$$

where  $\theta_1$  is the proficiency or latent ability level of an examinee on dimension 1;  $\theta_2$  is the proficiency or latent ability level of an examinee on dimension 2;  $e$  is the natural log base ( $e = 2.718$ );  $a_1$  and  $a_2$  are the item discrimination parameters on dimension 1 and dimension 2, respectively.  $d$  is related to item difficulty or the location of the item response surface.  $c$  is the guessing parameter or lower asymptote for the item.  $D$  is a constant,  $D=1.702$ .  $P_i(\theta|a,b,c)$  is the probability of an examinee with  $\theta_1$  and  $\theta_2$  two dimension proficiency scores responding correctly to an item.

As mentioned above, Lockwood, *et al.* (2007) and Martineau (2006) found that there was a violation of measurement invariance in a vertically scaled developmental score. The results of these studies showed that there was a shift in the construct across wide grade ranges, *i.e.*, there were violations of measuring similar content and the same weight of each main topic across grades. In this study, the item responses based on the two-dimensional 3P logistic IRT model was used to mimic real tests in vertical scaling situation.

In this study,  $\theta_1$  was viewed as a dominant ability or primary ability across all items and was based on an examinee's learning trait (DeMars, 2003). This general learning trait was of interest and was put on one common scale across grades.  $\theta_2$  was treated as a secondary or minor ability such as prior knowledge or the ability depending on a specific subtopic or a construct shift across tests. In general, the conditional probability of an examinee's correct response on the common item depended on both the dominant ability and the secondary ability because the common items are off-grade items. However, only the dominant ability was of interest in this vertical scaling. The secondary or minor ability was viewed as a kind of contamination of the dominant or general ability.



The unidimensional 3P logistic IRT model (refer to Eq.1) was then used for conducting vertical scaling across grades. The 3P logistic model was selected because it is one of the most popular IRT models in state assessment programs such as those in California, Colorado, and Florida. Secondly, the guessing parameter,  $c$ , is important to estimates for students with lower ability levels. If the guessing parameter is not included, the estimates of extreme low ability could be a problem.

### **3.1.3 Manipulated factors**

As discussed in Chapter 2, a number of technical factors affect the vertical scaling results. In this study, three main technical factors were manipulated while the other factors were fixed. One was the degree of violation of multidimensionality, which was achieved by adjusting the correlations between two latent abilities. Calibration and scaling methods were also manipulated factors. In addition, the variances of latent ability distributions for the six grades were also manipulated in this study. The fixed factors included sample size of each grade, ability estimation method, number of common items, and total number of items in each grade.

#### **3.1.3.1 Intertrait Correlation**

One fundamental assumption of using the logistic IRT model is unidimensionality. In the vertical scaling context, it is assumed that item responses are based on the same skill or same composite of multiple skills across all grade levels. If the assumption of unidimensionality is met, students' performances on grade level items and common items, which are used for adjacent grades, should be consistent. Under these

circumstances, only one general ability or dominant ability,  $\theta_1$ , contributes to students' performances. The unidimensional condition is the desired test condition and was used as a reference in this study to compare the scaling results from the conditions that violated the assumptions.

For the specific below-grade/on-grade items design, the common items were from the lower adjacent grade level test. The students' performances on on-grade level items and common items are likely inconsistent. So, the assumption of unidimensionality is not met. A within-item multidimensional structure was assumed in this study, *i.e.*, both a dominant ability,  $\theta_1$ , and a secondary ability,  $\theta_2$ , contributed to students' performance on common items.  $\theta_1$  is a dominant ability across all items and grades and is based on examinee's learning trait (DeMars, 2003),  $\theta_2$  is a secondary or minor ability such as prior knowledge or an ability that depends on a specific subtopic or a construct shift across tests. It exhibited when students work on common items.

The two ability distributions,  $\theta_1$  and  $\theta_2$ , of each grade were assumed to be normally distributed with the same means and fixed variance. The joint distribution of examinees' abilities is a multivariate normal distribution,  $MVN_2(\mu, \Sigma)$  (it is also a bivariate normal distribution because there are only two latent abilities), where  $\mu$  is the vector of the means of two latent abilities,  $\theta_1$  and  $\theta_2$ . The  $\Sigma$  is the variance-covariance matrix of two latent abilities,  $\theta_1$  and  $\theta_2$ .

The  $\Sigma$  for each grade is:  $\begin{pmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ . Both  $\sigma_{12}$  and  $\sigma_{21}$  are the covariances between two latent abilities,  $\theta_1$  and  $\theta_2$ .  $\sigma_1^2$  and  $\sigma_2^2$  are the variances of two latent abilities,  $\theta_1$  and  $\theta_2$ , respectively. In this study,  $\sigma_1$  and  $\sigma_2$  were assumed to be same.

The covariance between  $\theta_1$  and  $\theta_2$ ,  $\sigma_{12}$ , was represented by the correlation between  $\theta_1$  and  $\theta_2$ .

To investigate the robustness of different calibration and scaling methods to the degree of violation of the unidimensional assumption, the correlations  $r_{12}$  between two abilities,  $\theta_1$  and  $\theta_2$ , were manipulated.

In some MIRT related studies, the correlations between two latent abilities have a wide range, 0 to 0.9 (Finch, 2011). However, in a vertical scaling context, it is not very likely that there is almost no correlation between two latent abilities. When the correlation between two latent abilities is very high such as 0.9, the data structure can be viewed as essentially unidimensional.

In this study, the two latent traits were simulated to be correlated at 0.6 and 0.3. 0.3 reflected the low correlation between the dominant ability distribution and the secondary ability distribution which is viewed as the contamination of the primary ability. 0.6 was the medium correlation between two latent abilities.

### **3.1.3.2 Calibration and scaling methods**

Concurrent calibration is one of the most commonly used IRT calibration methods. When unidimensionality holds and data from a few grade levels are calibrated, the concurrent calibration is efficient and provides adequate estimation (Hanson & Béguin, 2002; Kim & Cohen, 1998). When unidimensionality does not hold and data from a wide range of grades are calibrated, the behavior of concurrent calibration is suspicious. In these situations, separate calibrations are usually thought to be more robust since it breaks the wide range into smaller units for scaling. However, very limited research has

been conducted to investigate these calibration and scaling methods in the vertical scaling context, especially under violation of unidimensionality assumptions.

In this study, five calibration and scaling methods were explored and compared: concurrent calibration, separate calibration by using mean/sigma transformation, separate calibration by using TCC transformation, pair-wise calibration, and semi-concurrent calibration by using ability transformation.

*Concurrent calibration.* When using concurrent calibration, a common scale was established by calibrating data from the six tests in a single computer run. In MULTILOG, concurrent calibration with multiple groups method was used at this situation. The common items between adjacent grade levels are used as a bridge to link adjacent grade levels. An examinee's item response pattern included item response data for the items taken and a "not reached" code was used for the items not taken.

For the 15 common items condition, it assumed that each student took 225 items total (6 grades x 50 total items, and 5 common-item sets, each with 15 items, need to be deducted since they are counted twice, so the total number is 225). In other words, it assumed that grade 3 students took items 1 to 50; and items 51 to 252 were not reached. For grade 4 students, each student took items 36 to 85, but not items 1 to 35 and 86 to 225. Items 35 to 50 are the common items between grade 3 and grade 4 which were taken by both grades. For grade 5 students, each student took items 71 to 120. The items 1 to 70 and items 121 to 225 were treated as "not reached". For grade 6 students, items 106 to 155 were answered and the other were not reached. For grade 7 students, items 141 to 190 were answered and the other were not reached. For grade 8 student, only the last 50 items, items 176 to 225, were answered. After the concurrent calibration, the item parameter estimates and ability estimates for all grades were on the same scale.

Typically, scale transformations were not necessary under this procedure. However, to compare the ability estimates with the “true” abilities, the overall scale will be adjusted so that the distribution of grade 5 ability estimates has a mean 0 and a variance 1 because grade 5 is the base grade with a distribution  $N(0,1)$ . That is, each student’s ability was subtracted by the mean ability estimates of grade 5 and then divided by the standard deviation of grade 5 ability estimates.

*Separate calibration with different scaling methods.* Different from concurrent calibration, separate calibration was accomplished in two main steps: 1) a separate calibration by using IRT estimation software and 2) statistical grade-by-grade chained scaling or linking. In the first step, the six grades were calibrated separately to obtain item parameter estimates for each grade. Six computer runs were needed in this step. After separate calibration, the item parameter estimates for each grade were on different metric due to the IRT model indeterminacy.

In the second step, the scale of grade 5 was selected as a base scale. Then a series of statistical transformation methods were used to linearly transform the other scales to the grade 5 scale to achieve a common scale. The series of statistical methods were based on two sets of the estimated item parameters, which were obtained from two separate calibrations for the common items between adjacent grades. As discussed in Chapter II, different statistics of the item parameter estimates were then used to get different transformation parameters, A and B.

The Mean/sigma method (Marco, 1977) uses the information of b-parameter estimates from the common items to obtain estimates of A and B. The TCC method (Stocking & Lord, 1983) minimizes the difference between the test characteristic curves of common items based on two separate calibrations.

The attractive feature of the mean/sigma method is statistical simplicity. The limitation of this method is that not all available information from the item parameters is used simultaneously. The salient feature of the TCC method is that all available information from item parameters (a, b, and c parameters) is used in the estimations. Theoretically, the more information included, the better are the estimates. However, the characteristic curve method requires complex iterative multivariate searching to obtain linking parameter estimates.

In practice, mean/sigma and TCC are the most popular scale transformation methods. Several studies have compared these methods under different equating conditions (Baker & Al-Karni, 1991; Hanson & Béguin, 2002; Lee & Ban, 2010). The comparison studies suggest that characteristics curves method may produce slightly better results than the mean/sigma method.

In the vertical scaling design, choosing and implementing scale transformations requires more attention because there is typically a mismatch of one group of examinees and the difficulty of the commons items. Hanson and Béguin (2002) suggested that it would be beneficial to apply multiple linking procedures and compare the scaling results. This would help in understanding the various issues and aspects of the scaling situation and in choosing the most appropriate scaling method. Therefore, two popular transformation methods, mean/sigma and TCC were investigated and compared in this study.

For grade 4 and grade 6, the item parameter estimates were transformed to the grade 5 base scale by using transformation coefficients directly. For the other grades, such as grade 3, grade 7, and grade 8, cumulative linking (or chained linking, FCAT, 2001) was used. For example, to transform the estimates of grade 3, they were

transformed on the grade 4 scale first. Then the transformed estimates on the grade 4 scale were transformed on the grade 5 scale. To transform the estimates of grade 7, the estimates were transformed on the grade 6 scale first. Then the transformed estimates on the grade 6 scale were transformed on the grade 5 scale. This cumulative linking is repeated for grade 8 until all grades were on the common base scale.

Using grade 5 as the base scale, the cumulative linking equations for grade 3, grade 4, grade 6, grade 7, and grade 8 are:

$$Y_{3 \rightarrow 5} = a_{34}a_{45}X_3 + a_{45}b_{34} + b_{45} \quad (17)$$

$$Y_{4 \rightarrow 5} = a_{45}X_4 + b_{45} \quad (18)$$

$$Y_{6 \rightarrow 5} = a_{65}X_6 + b_{65} \quad (19)$$

$$Y_{7 \rightarrow 5} = a_{76}a_{65}X_7 + a_{65}b_{76} + b_{65} \quad (20)$$

$$Y_{8 \rightarrow 5} = a_{87}a_{76}a_{65}X_8 + a_{76}a_{65}b_{87} + a_{65}b_{76} + b_{65} \quad (21)$$

where  $a_{ij}$  and  $b_{ij}$  are the coefficients of transformation parameters, A and B, between adjacent grades respectively.  $X'_i$ s are the ability parameter estimates based on the separate calibrations.  $Y$ s are the transformed ability estimates on the common scale.

*Pair-wise calibration with scaling method.* Pair-wise calibration is a hybrid calibration method in vertical scaling (Karkee, *et. al.*, 2003). It combines concurrent and separate calibration methods and was accomplished in two steps. In the first step, each pair of adjacent grades was concurrently calibrated to obtain item parameter estimates. Grades 3 and 4 were concurrently calibrated. Similarly, the items in grades 5 and 6 were concurrently estimated, as were the items in grades 7 and 8.

In the second step, the item parameter estimates of grades 5 and 6 were used as the base scale. The item parameter estimates of grades 3, 4, 7, and 8 were then transformed on this scale. To transform the item parameter estimates of grades 3, 4, 7,

and 8 on a common scale, the TCC linking method (Stocking & Lord, 1983) was used for the transformation. For grades 3 and 4, the TCC linking parameters were computed from the common items between grade 4 and 5. For grades 7 and 8, the TCC linking parameters were computed from the common items between grade 6 and 7.

*Semi-concurrent calibration with ability transformation.* Semi-concurrent calibration is a relatively new proposed calibration method in vertical scaling (Meng, 2007). It is also a hybrid method that is accomplished in two steps. In the first step, grades 3, 4, and 5 were concurrently calibrated with a multiple groups method to obtain item parameter and ability estimates for these grades; grades 5, 6, 7, and 8 were calibrated together to get item parameter and ability estimates. These estimates for items and abilities from the two calibrations were on different scales due to the indeterminacy problem. Therefore, this yielded two sets of ability estimates for grade 5 and they are on different scales.

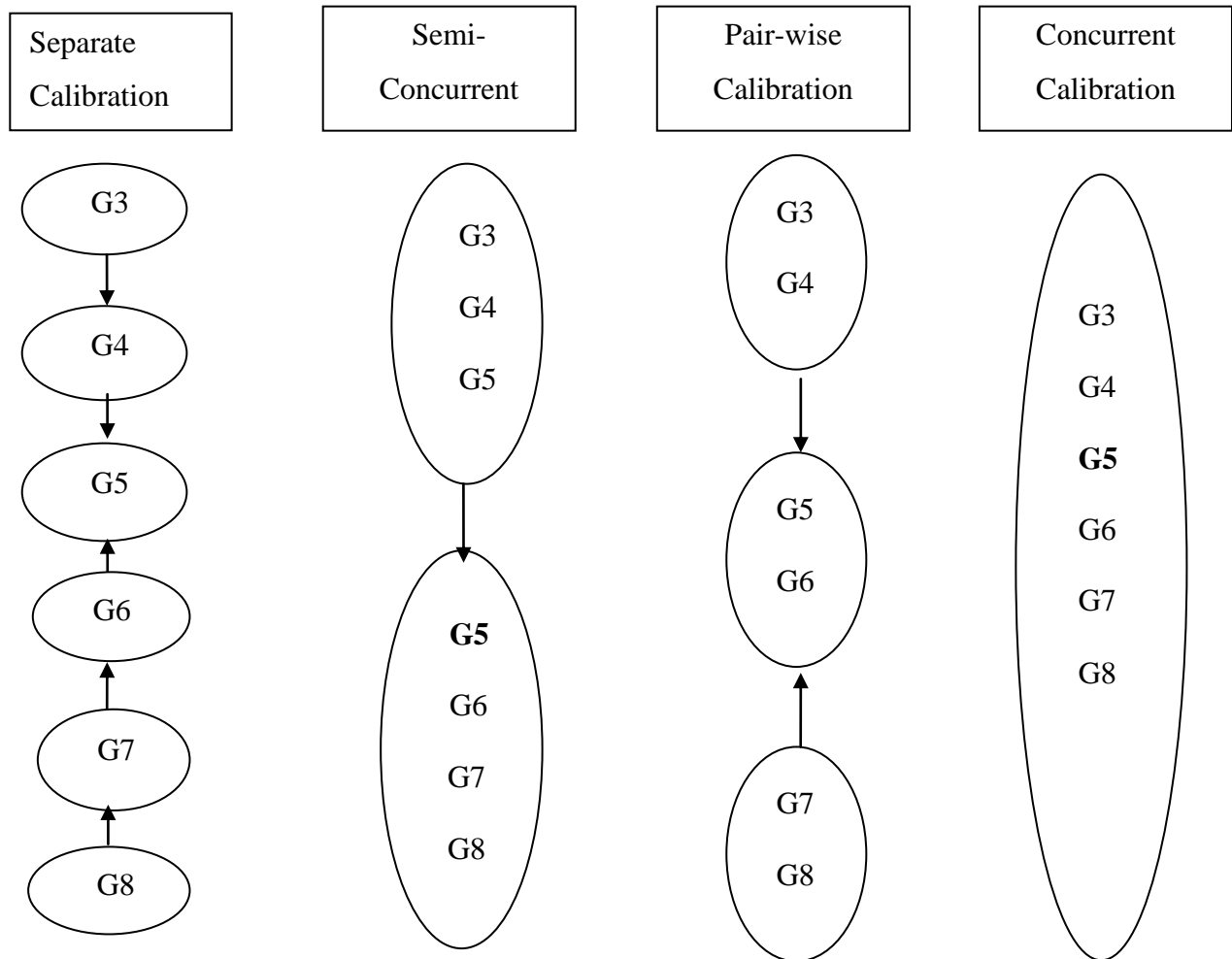
In the second step, the scale for grades 5, 6, 7, and 8, were used as the base scale. The first set of item parameter estimates (grades 3, 4, and 5) were linearly transformed to the second set of item parameter estimates (grades 5, 6, 7, and 8) by using proficiency transformations (also referred as ability transformations). The transformation parameters, A and B, were obtained by using the mean and standard deviation of the two sets of grade 5 ability estimates:  $(\theta^* - \bar{X}_{52})/S_{52} = (\theta - \bar{X}_{51})/S_{51}$ . Because the second set of estimates of grade 5 were specified,  $S_{52} = 1$  and  $\bar{X}_{52} = 0$ , the grade 3 and 4 estimates were transformed on the base scale by:  $\theta^* = (1/S_{51})\theta - (1/S_{51})\bar{X}_{51}$ , where  $S_{51}$  and  $\bar{X}_{51}$  are the standard deviation and mean, respectively, of grade 5 ability estimates in the first calibration. For example, to put grade 4 ability estimates on the base scale, grades 5, 6, 7, and 8 scale, all grade 4 ability estimates were



divided by the standard deviation of grade 5 ability estimates and then were subtracted by the weighted mean of grade 5 ability estimates yielded in the first half of the concurrent calibration.

As described above, both pair-wise calibration and semi-concurrent calibration are hybrid methods of the concurrent method and the separate calibration method. They are expected to improve the scaling results since it utilizes the desired features of both concurrent calibration and separate calibrations. They break the wide range of grades into narrower parts and still keep some attractive features of concurrent calibration. They also use direct transformations instead of long-chained transformations for the grades further away from base grade 5 which may reduce the transformation errors.

In summary, four calibration methods were investigated in this study. For the separate calibration, two popular scaling methods were used for linking item parameter estimates. So, a total of five calibration and scaling methods were investigated in this study. A graphical representation of the four calibration methods is provided in Figure 3.2:



**Figure 3.2** Graphical representation of the separate, semi-concurrent, pair-wise, and concurrent calibrations

### 3.1.3.3 The variances of ability distributions for 6 grades

In this study, MULTILOG was used for item parameter and latent ability estimation. The variance or standard deviation of latent ability distribution is usually set as 1 by MULTILOG defaults. However, in practice, the variances of student test score distributions are not always equal across grades. The variances likely increase from lower to higher grades as is the case within Iowa Test Batteries (Kolen & Brennan, 2004; Tong & Kolen, 2007). Therefore, the variances (or standard deviations) of ability

distributions for each grade were also manipulated in this study to examine how the vertical scales were affected by the different variances of ability distributions.

There were two levels for this factor. One is that the standard deviations of ability distributions were all fixed at 1 for grade 3 to grade 8. Another one is that the standard deviations of ability distributions are assumed as 0.9, 0.9, 1, 1, 1.1, 1.1 for grade 3 to grade 8, respectively. The standard deviation of grade 5 and 6 were set at 1 because they are the middle grade. The standard deviations of ability distribution for grade 3 and 4 were assumed the same and slightly lower than grade 5. The standard deviations of ability distribution for grade 7 and 8 are the same, but higher than the lower grades as seen in practice (Kolen & Brennan, 2004). This factor was investigated to examine how the variance of ability distributions affects the vertical scaling results.

#### **3.1.3.4 Fixed factors**

Theoretically, the larger the sample size, the more accurate the scaling results. If the sample size is too small, the item parameter estimates are unstable. However, in practice, a larger sample size may be difficult to obtain. In this study, the number examinees in each grade was fixed at 2000 since the literature suggests that sample sizes of at least 1000 can produce adequate estimates for the 3P logistic model (Fitzpatrick & Yen, 2001; Liu & Walker, 2007). Therefore, across all six grades, 12,000 examinees were simulated.

The number of common items was also fixed in this study. Based on the literature (Kolen & Brennan, 2004), 20% of the total items is the minimum number of common recommended in practice. Typically, the larger the common item set, the smaller the mean squared error. However, when the total test length is fixed, if common

items increase, the grade level items decrease. If there is a violation of unidimensionality, it leads to more discrepancy between the performances on common items and grade level items. This discrepancy may further affect the robustness of the scaling methods. Therefore, 30% of the total items is a reasonable number for common items. Since the number of total items in each grade was fixed at 50, the scaling design with 15 common items was used in this study.

MAP ability estimation was used to obtain the examinees' ability estimates. A few studies have been conducted to compare how Bayesian methods, MAP and EAP, and MLE behave in vertical scaling contexts (Tong & Kolen, 2007; Briggs & Weeks, 2009). In general, Bayesian methods produce slightly better results than the MLE method. One possible reason is that MLE ability estimates become very unstable in the tails of the ability distribution. In addition, the arbitrary specification for the very extreme response patterns may distort the vertical scales. Therefore, MAP, one of the two Bayesian methods, was used to obtain the ability estimates in this study.

#### **3.1.4 Number of replications**

Usually, more replications produce less biased parameter estimates when empirical sampling distributions are explored. When comparing IRT-based methodologies, empirical sampling distributions are not necessary and a small number of replications may be sufficient (Harwell *et. al.*, 1996). Since this simulation study will compare different IRT-based vertical scaling methods rather than explore the sampling distribution of estimated ability means or other statistics, the number of replications was set at 200.

### 3.2 EVALUATION AND COMPARISON CRITERIA

Because no generally accepted growth definition or growth model exists in the literature, the characteristics of the vertical scales cannot be compared relative to an absolute criterion (Tong & Kolen, 2007). However, the scaling results can be compared with themselves. For simulated data, because the true value of ability is already known, the scaling results can be compared with the “true” scale to evaluate how the different methods behave. In this study, both bias and root-mean-squared-deviation of the ability estimates were used as criterion. The difference between the ability estimates and the ability parameter (only the dominant ability is of interest) was used to explore the direction of the bias. The mean bias of ability estimates for each grade was computed as:

$$\text{Mean bias of ability} = \frac{\sum(\hat{\theta}_i - \theta_i)}{2000} \quad (22)$$

where  $\theta_i$  is the “true” latent ability which is the defined parameter for data generation.  $\hat{\theta}_i$  is the estimate of “true” ability. The average bias across 200 replications was used as a criterion. The bias can be either negative or positive number which indicates the direction of the bias.

The root-mean-square-deviation (RMSD) of the ability estimates was used to evaluate the magnitude of the bias. The root-mean-square-deviation (RMSD) is expressed as the sum of the squared difference between the ability estimates and the ability parameter weighted by the total number of examinees in each grade (Stone, 2009):

$$RMSD = \sqrt{\frac{\sum_{i=1}^{2000} (\hat{\theta}_i - \theta_i)^2}{2000}} \quad (23)$$

The average RMSD across 200 replications was used as the final criterion:

$$\text{Average RMSD} = (\sum_{i=1}^{200} \text{RMSD}) / 200 \quad (24)$$

In addition, the correlations of the ability estimates and “true” abilities,  $r_{\hat{\theta}_i \theta_i}$ , were also used as a criterion to evaluate the different estimation methods. If the correlation is high, it means there is less change of the rank of students’ latent ability even if the variance of the whole distribution has changed.

A total of 30 (5 x 3 x 2) combinations of vertical scales were examined and compared in this study. In addition, the results of each grade level were compared within each vertical scale. Since there were six grade levels in each combination, 180 (30 x 6) averaged bias and RMSDs were compared in this study.

### 3.3 DATA GENERATION

#### 3.3.1 Software for data generation, calibrations, and scaling

The examinees’ responses were simulated by using proc IML in SAS 9.2 program. The item parameters and examinees’ two latent abilities,  $\theta_1$  and  $\theta_2$ , were simulated from the specified distributions.

After item responses were simulated, item parameters and latent abilities were then estimated using MULTILOG (2003). One desired feature of MULTILOG is that it

has the ability to handle multiple groups to perform a concurrent calibration with nonequivalent groups. For the concurrent calibration, the distribution of the first group, grade 3, is defined with mean 0 and variance 1 by MULTILOG. The common scale was transformed by using direct statistical procedures so that the middle grade 5 had a mean of 0 and variance of 1.

For the separate calibrations, additional transformations were needed to put separate estimates for each grade on one common scale. The computer program, ST (Hanson & Zeng, 2004), was used to obtain the scale transformation coefficients. ST is a program for computing the coefficients of IRT scale transformations. The basic IRT model in ST is the 3P logistic model. Four pieces of information for common items were needed in ST to obtain transformation coefficients for mean/sigma and TCC (Stocking-Lord) linking methods. The four pieces of information for common items included: item parameter estimates of the common items which were used as the base; item parameter estimates of the common items to be scaled; ability estimates based on the base metric; and ability estimates based on the scale which need to be transformed. All this information was written into a fixed format in ST. After the transformation parameters,  $A$  and  $B$ , were obtained from ST, the ability estimates of grade 3, 4, 6, 7, and 8 were then put on a common metric by using equations 17 to 21.

### **3.3.2 Data generation**

For this study, vertical scales were constructed spanning from grade 3 to grade 8, with the common-items nonequivalent groups design used as the linking design. In common-

items design, the examinees for each grade took both grade level items and a set of common items which were constructed for the adjacent grades.

The number of items for each test was fixed at 50 and the number of common items was 15. When the number of common items was 15, the number of grade level items was 35. To compare different calibration and scaling methods, grade 5 was used as the base scale. Because the same set of common items was used for adjacent grades, the item responses in all six grades were simulated together. For example, when the number of common items was 15 between adjacent grades, a total of  $225 \times 2000 \times 6$  (225 items, 2000 examinees, and 6 grades) item responses were generated simultaneously. The item responses of the 50 items appropriate for that grade were coded as 1 or 0. The other 175 items were treated as “non-reached” and coded as 9.

The main steps for simulating item response were:

- 1) *Item parameter.* The item discriminations of the first dimension,  $a_1$  (refer to Eq. 16) were from a uniform distribution in the range of  $[0.75, 1.5]$ . The item discriminations of the second dimension,  $a_2$ , were assumed to be different. The specification of  $a_2$  was based on whether the items were common items or non-common items. For non-common items, the item discriminations of the secondary dimension are fixed as 0 to reflect unidimensionality of these items. When students worked on common items, the item discriminations for the secondary dimension were fixed at 0.4. The  $a_2 = 0.4$  was selected because it is a minor dimension so that the ratio of the discrimination of the minor dimension to dominant dimension,  $a_2/a_1$ , mostly falls between  $1/3$  and  $1/2$ . According to previous studies (Finch, 2011; Zhang & Stone, 2008), for an approximate simple



structure multidimensional test, the discrimination of the minor dimension is always smaller than the discrimination of the dominant dimension. That is, only the dominant ability contributes to the item responses for non-common items because the secondary ability is cancelled out by the 0 discriminations for the secondary dimension. Both dominant ability and secondary ability contribute to the item responses for common items.

In general, the data structure is unidimensional or essentially unidimensional for non-common items. The data structure is multidimensional structure or non-simple structure for common items. That is, there was one set of item parameters across the adjacent two grades for the common items. This data structure reflected the multidimensionality because of the construct shift from students in the either lower grade responding as well as students in higher grade responding. For example, the structure of items 1 to 35 was unidimensional. The structure of common items between grade 3 and grade 4, items 36 to 50, was multidimensional. The structure of items 51 to 70 was unidimensional. The structure of items 71 to items 85 was then multidimensional because they were the common items between grades 4 and 5 (Figure 3.3).

The  $d$  parameters (refer to Eq. 16) for each grade were from a normal distribution (Zhang & Stone, 2008) but with different means and the same standard deviations for different grades since item responses across all grades were generated simultaneously. The distribution of  $-d$  were  $(-1.0, 1)$ ,  $(-0.45, 1)$ ,  $(0, 1)$ ,  $(0.35, 1)$ ,  $(0.6, 1)$ , and  $(0.8, 1)$ , for grade 3 to grade 8, respectively. The increased means indicate that the items were more difficult from grade 3 to grade 8. The corresponding mean of each grade ability distribution,  $-1.0$ ,  $-0.45$ ,  $0$ ,  $0.35$ ,

0.6, and 0.8 for grade 3 to grade 8, respectively, was added to the standard normal distribution (0, 1) so that  $d$  parameters were concentrated the center of corresponding ability distribution. The  $c$  parameter was fixed as 0.2 for all items to correspond to a random guessing model for multiple choice items with 5 options (DeMars, 2006).

Grades	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Grade 3	G3_on	G3_4c										
Grade 4		G3_4c	G4_on	G4_5c								
<b>Grade 5</b>				G4_5c	G5_on	G5_6c						
Grade 6					G5_6c	G6_on	G6_7c					
Grade 7							G6_7c	G7_on	G7_8c			
Grade 8										G7_8c	G8_on	

**Figure 3.3** Simulated multidimensional data structure

(Note: The rectangles with green indicate the items are unidimensional. The rectangles with grey indicate the items are multidimensional.)

- 2) *Ability distributions.* Two ability distributions, the dominant ability  $\theta_1$  and the secondary or minor ability  $\theta_2$ , were simulated for 2000 examinees from each grade.  $\theta_1$  and  $\theta_2$  were from a multivariate normal distribution with the same means and same variances, and different correlation levels. The means of the

distributions of two abilities,  $\theta_1$  and  $\theta_2$ , for grade 3 to grade 8 are: (-1.0, -1.0), (-0.45, -0.45), (0, 0), (0.35, 0.35), (0.6, 0.6), and (0.8, 0.8), respectively. The means of the distributions were assumed from a common scale with patterns of decelerating growth across grade levels (FCAT, 2006; Tong & Kolen, 2006; Briggs & Week, 2009). The variances of the two abilities for each grade are all fixed at 1 or with different variances for six grades. The correlations between two latent abilities were manipulated with 2 levels: 0.3 and 0.6.

- 3) *Computing probability of correct responses.* The simulated item parameters and simulated abilities were used to calculate the probability for correctly answering each item. The probability of correct responses is calculated by putting the simulated item parameters and abilities in the 2-dimensional 3P logistic MIRT model (refer to Eq.16).
- 4) *Comparing probability to a random number.* A random number from a uniform distribution  $U(0,1)$  was generated and compared to the computed probability of a correct response. If the random number was greater than the probability, the response was 0; otherwise, the response was 1 (Stone, 2009; Finch, 2011).
- 5) *Repeating steps.* Steps 1 and 4 were repeated 200 times.

### 3.3.3 Validating data generation

To validate data generation, the statistics (means and variances) of generated latent abilities were checked by using *proc means* in SAS. The correlations between the two simulated ability distributions for each grade were checked with the *proc corr*

procedures in SAS. In general, the statistics and correlations were very close to the specified parameters.

The multidimensional data structure was checked by software program NOHARM (Fraser & McDonald, 1988) and Mplus (Muthen & Muthen, Version 6.11). The simulated item responses were checked with exploratory factor analysis (EFA) with one and two dimensions in both NOHARM and Mplus.

First, in NOHARM, the model fit was checked by comparing the sum of squares of residuals, root mean square of residual and Tanaka index of goodness of fit. Overall, the smaller sum of squares of residuals and root mean square of residual, and the larger Tanaka index of goodness of fit, suggests that the two dimensional model fits the simulated data better than the one dimensional model. For example, under the condition with a correlation between the two latent abilities of 0.3 for students in a grade, when one dimension EFA was used, the sum of squares of residuals =0.041; root mean square of residual= 0.006; and the Tanaka index of goodness of fit = 0.988. If a two dimensional model was used, sum of squares of residuals =0.031; root mean square of residual= 0.005; and Tanaka index of goodness of fit = 0.991. These indices suggest that the two dimensional model fits the item responses data set better than the unidimensional model.

The factor loadings were examined using Mplus. Within a grade, the data structure of the test (including unidimensional items and multidimensional items) is multidimensional. For example, under the condition with 10 common items and a correlation between two latent abilities of 0.3 for a grade test, based on the Geomin rotated loadings, it showed that all items have obvious larger factors loadings on factor 1. Most factor loadings were from 0.4 to 0.7. The common items (the first 10 items in

this example) had small factors loadings on factor 2, with most loadings from 0.2 to 0.3. However, factor loadings were very small for non-common items (items 11 to item 50 in this example) and most were close to 0 (refer to Appendix A table 1). This factor loading pattern matched the expected data structure. The factor structure table (Appendix A table 2) also verified this multidimensional data structure. When the items were common items, they had within-item multidimensionality. When the items were non-common items, they were essentially unidimensional. In conclusion, these examinations suggest the data generation was valid.

### **3.3.4 Implementation of the study design**

The procedures for implementing the simulation study for each calibration method were:

#### **1. Concurrent calibration**

- 1) Simulated examinees' item responses. The item responses were generated based on the above data generation procedures.
- 2) Calibrated item responses concurrently. Ran MULTILOG one time with 6 grades together to obtain item parameter estimates.
- 3) Ran MULTILOG again to score 6 grades together.
- 4) Filtered MULTILOG results from step 3 to get ability estimates.
- 5) Computed bias and RMSD between the true ability and ability estimates.
- 6). Repeated step 1-5 200 times.
- 7). Summarized and analyzed the results across 200 replications.

#### **2. Semi-concurrent (or Hybrid) calibration**

- 1) Simulated examinees' item responses. The simulated item responses were the same as in the concurrent calibration step 1.
- 2) Calibrated item responses concurrently for grades 3, 4, and 5. First ran MULTILOG to obtain item parameter estimates with grades 3, 4, and 5 together; then ran MULTILOG to get item parameter estimates with grades 5, 6, 7, and 8 together;
- 3) Ran MULTILOG again to obtain ability estimates for grades 3, 4, and 5 concurrently; ran MULTILOG again to obtain ability estimates for grades 5, 6, 7, and 8 concurrently.
- 4) Filtered MULTILOG results from step 3 to get ability estimates.
- 5) Using the proficiency transformation method (PT), two separate ability estimates were put on the scale of grades 5, 6, 7, and 8.
- 6) Computes bias and RMSD between the true ability and ability estimates.
- 7) Repeated step 1-6 200 times.
- 8) Summarized and analyzed the results across 200 replications.

### 3. Pair-wise calibration

- 1) Simulated examinees' item responses. The simulated item responses were the same as in the concurrent calibration step 1.
- 2) Calibrated item responses concurrently for each paired adjacent grades, grades 3 and 4, grades 5 and 6, and grades 7 and 8, respectively. First ran MULTILOG to obtain item parameter estimates with grades 3 and 4; then ran MULTILOG to obtain item parameter estimates with grades 5 and 6; the same procedure for grades 7 and 8.

- 3) Ran MULTILOG again to get ability estimates for grades 3 and 4 concurrently; ran MULTILOG again to get ability estimates for grades 5 and 6; then for grades 7 and 8.
- 4) Filtered MULTILOG results from step 2 and 3 to obtain item parameter estimates and ability estimates.
- 5) Used ST program to obtain transformation coefficients for TCC method.
- 6) Transformed the ability estimates for grade 3, 4, 6, 7, and 8 on the grade 5 scale by using linking coefficients which were obtained from step 5.
- 7) Computed bias and RMSD between the true ability and ability estimates.
- 8) Repeated step 1-7 200 times.
- 9) Summarized and analyzed the results across 200 replications.

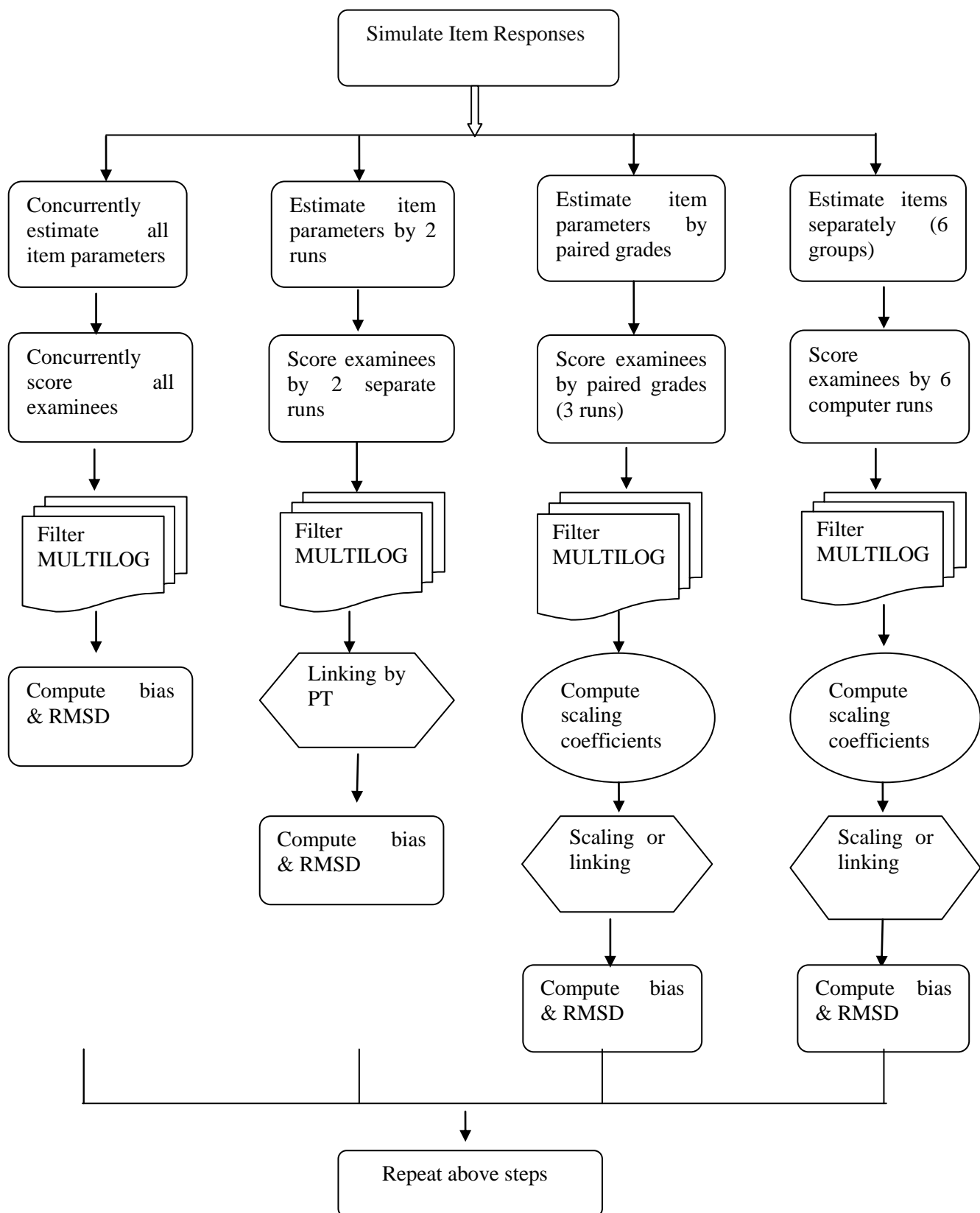
#### 4. Separate calibration

- 1) Simulated examinees' item responses. The simulated item responses were the same as in the concurrent calibration step 1.
- 2) Calibrated item responses separately for 6 grades. First ran MULTILOG 6 times for each grade to get item parameter estimates;
- 3) Ran MULTILOG 6 times again to obtain ability estimates of 6 grades separately.
- 4) Filtered MULTILOG from step 2 and step 3 to obtain each set of common item parameter estimates and ability estimates.
- 5) Uses ST program to obtain transformation coefficients for mean/sigma and TCC method.
- 6) Transformed the ability estimates for grades 3, 4, 6, 7, and 8 on the grade 5 ability's scale by using linking coefficients which were obtained from step 5.

- 7) Computed bias and RMSD between the true ability and ability estimates.
- 8) Repeated step 1-7 200 times.
- 9) Summarized and analyzed the results across 200 replications.

In addition, the procedures of implementing the simulation study are also illustrated by the following flow chart (figure 3.4):





**Figure 3.4** Flow chart of the simulation study implementation

## 4.0 RESULTS

In this chapter, the results from the simulation study are presented. The five calibration and linking methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma, and separate with SL linking, are compared at different test conditions. There are three main sections in this chapter: comparisons of calibration and scaling methods under unidimensional condition; comparisons of calibration and scaling methods under multidimensional conditions with  $r=0.3$  between two latent abilities; and comparisons of calibration and scaling methods under multidimensional conditions with  $r=0.6$  between two latent abilities.

In each section, the average bias, RMSD, and the correlation between ability estimates and “true” abilities are reported for each test condition. At last, the study results are summarized under each condition.

As mentioned in Chapter 3, the difference between the ability estimates and the ability parameter (only the dominant ability is of interest) is called bias. The average bias across 200 replications was used as a criterion for comparison. The root-mean-square-deviation (RMSD) of the ability estimates is the sum of the squared difference between the ability estimates and the ability parameter weighted by the total number of examinees in each grade. The average RMSD across 200 replications was used to evaluate the magnitude of the bias. In addition, the correlation of the ability estimations

and “true” abilities was also used to evaluate the different estimation methods as it provides some information about the change on the rank of examinees’ ability.

## 4.1 COMPARISONS OF SCALING METHODS UNDER UNIDIMENSIONAL CONDITION

### 4.1.1 Bias of ability estimates

The bias results of the unidimensional condition with fixed variance of 1 and varied variances of distributions are reported in Tables 4.1 and 4.2, respectively. The unidimensional condition is a desired condition and is used for reference and comparison. In each table, the average biases of ability estimates,  $\hat{\theta}_i - \theta_i$ , are compared for the five calibration and linking methods from grade 3 to grade 8. First, the mean bias of each grade with 2000 students was calculated. Next, the condition was replicated 200 times. Then the average bias was calculated based on the 200 mean biases. The value in each cell is the average bias across 200 replications. The value in the parenthesis is the standard deviation (SD) of the 200 mean biases across replications. Smaller SD values indicate less variability in estimates across replications. Figures 4.1 to 4.2 demonstrate the comparisons among the five calibration and linking methods.

Table 4.1 provides the average bias for each calibration and scaling method for the unidimensional condition with fixed variance 1. Most bias index values were very close to 0 or negative, indicating that the average estimated abilities were overall either close to “true” abilities or smaller than “true” abilities. The average biases under pairwise calibration and separate calibration with both Stocking-Lord (SL) linking and

Mean/Sigma (MS) linking were similar and relatively small for all six grades, with absolute values falling between 0 and 0.1 units of ability parameter. The biases were also very consistent across the six grades for these two methods. For the pair-wise calibration, the average biases were smaller than they were in the separate calibration. Most of them were close to 0. That is, there were very small average biases under the pair-wise calibration. Finally, Figure 4.1 graphs the average bias for the methods and displays flat and identical trends except for concurrent and semi-concurrent methods.

The average biases under concurrent calibration and semi-concurrent calibration for grade 3 to grade 6 were consistent and similar to the bias pattern of the separate calibration and pair-wise calibration. However, for grade 7 and 8, the average biases were negative with larger absolute values, especially under the concurrent calibration. Specifically, under the concurrent calibration, the average bias for grade 7 (about 0.3 units of ability parameter) was almost twice the size of the average bias for grade 6 (0.14 units) while the average bias for grade 8 almost tripled the average bias for grade 6. Under semi-concurrent calibration, the absolute average bias for grade 7 and grade 8 (-0.194 and -0.28 respectively) were also larger than those for the lower level grades (less than 0.1). These values indicated that the ability was underestimated generally for grade 7 and 8 under concurrent calibration and semi-concurrent calibration.

For the unidimensional condition with varied variances for the ability distributions across 6 grades, the bias patterns under all five calibration methods (Table 4.2 and Figure 4.2) were very close to the bias patterns for the unidimensional condition with fixed variance. The average biases under separate calibration with SL linking and MS linking and pair-wise calibration were also very small and consistent across the six

grades. For the concurrent calibration and semi-concurrent calibration, the overall bias patterns were similar. Most differences were between 0.001 to 0.01 units.

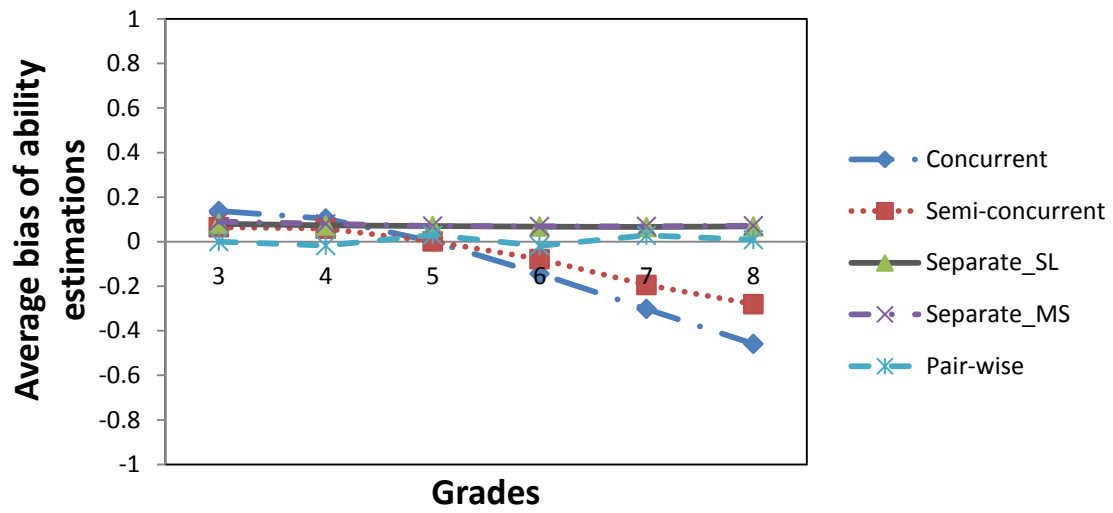
In general, under both fixed and varied variance conditions, the average biases under separate calibrations with SL linking and MS linking and pair-wise calibration were very small across grades, with absolute values below 0.1 units. The standard deviation (SD) of the average bias estimates for base grade 5 was the smallest, with values increasing as the grades become further from the base grade 5. The estimates of average bias were more consistent for the grades closer to the base grade than those further away from the base grade. For the lower grades, 3 to 6, the bias patterns were similar among all five methods. However, for the higher grades 7 and 8, the bias patterns under concurrent and semi-concurrent calibrations were different from the other three methods, with concurrent calibration producing the largest average bias. This indicates that these two methods, to some extent, underestimate the latent abilities of the students in the higher grades.

**Table 4.1** Average bias of  $\hat{\theta}$  under unidimensional condition with fixed variance

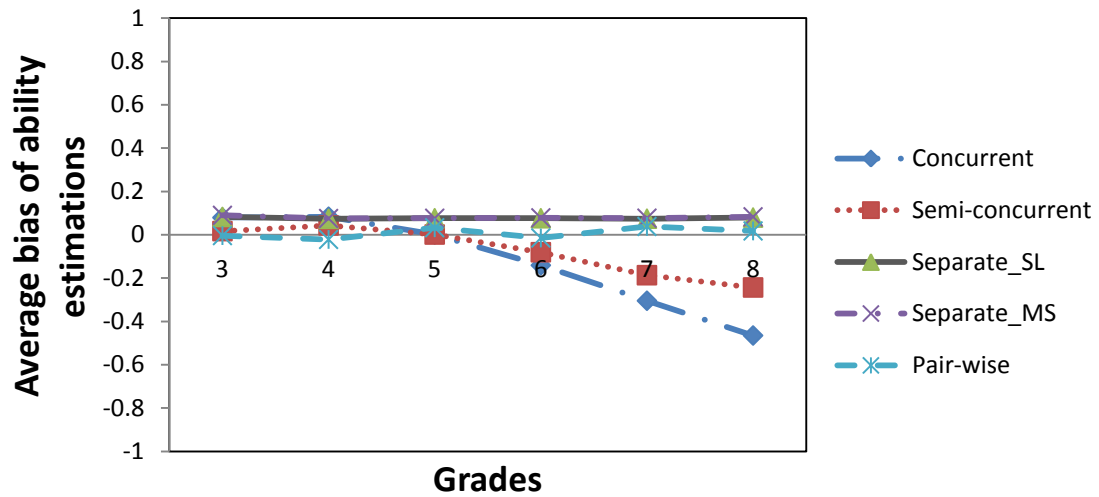
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.138(0.040)	0.065(0.034)	0.081(0.057)	0.091(0.097)	0.000(0.02)
Grade 4	0.104(0.031)	0.059(0.028)	0.073(0.044)	0.081(0.054)	-0.018(0.039)
Grade 5	0.001(0.022)	0.001(0.022)	0.071(0.033)	0.071(0.033)	0.030(0.025)
Grade 6	-0.144(0.030)	-0.078(0.091)	0.068(0.043)	0.068(0.052)	-0.019(0.032)
Grade 7	-0.302(0.037)	-0.194(0.148)	0.066(0.050)	0.069(0.068)	0.028(0.075)
Grade 8	-0.459(0.039)	-0.280(0.188)	0.068(0.057)	0.072(0.086)	0.009(0.078)

**Table 4.2** Average bias of  $\hat{\theta}$  under unidimensional condition with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.079(0.042)	0.017(0.037)	0.082(0.053)	0.089(0.091)	-0.004(0.048)
Grade 4	0.083(0.032)	0.043(0.028)	0.073(0.042)	0.075(0.054)	-0.022(0.036)
Grade 5	0.001(0.021)	0.001(0.021)	0.076(0.033)	0.076(0.033)	0.032(0.026)
Grade 6	-0.141(0.031)	-0.081(0.092)	0.076(0.039)	0.078(0.048)	-0.015(0.033)
Grade 7	-0.305(0.037)	-0.187(0.148)	0.074(0.047)	0.076(0.067)	0.038(0.064)
Grade 8	-0.465(0.041)	-0.244(0.154)	0.079(0.050)	0.083(0.085)	0.017(0.069)



**Figure 4.1** Average bias of  $\hat{\theta}$  under unidimensional condition with fixed variance



**Figure 4.2** Average bias of  $\hat{\theta}$  at unidimensional condition with varied variances

#### 4.1.2 RMSD of ability estimates

The RMSD for the unidimensional condition with fixed and varied variances for ability distributions are reported in Tables/Figures 4.3 and 4.4, respectively. In each table, the average RMSD of ability estimates were compared for the five calibration and linking methods from grade 3 to grade 8. The value in each cell is the average RMSD of each grade over 200 replications and the standard deviation (SD) of the RMSD estimates is in the parentheses as for the bias results. The small SD indicates small variability in RMSD estimates across replications.

Under the unidimensional condition with fixed variance, the performance of the five calibrations and linking methods were somewhat different across the six grades. For grade 3 to grade 6, the RMSDs for all five methods were comparable and small, with most values being about 0.3 units of ability parameters or slightly lower. For grade 7 and 8, the RMSDs under concurrent and semi-concurrent calibrations were higher, with the semi-concurrent calibration performing the worst for these grades as shown in Figure 4.3. The RMSD with concurrent calibration was about 0.537 for grade 8, which almost doubled the RMSD for the base grade 5 (0.269 units) under concurrent calibration. The RMSD with semi-concurrent calibration is about 0.89 for grade 8, which is almost three times that for most RMSDs yielded by the other three methods for all grades. A RMSD of 0.89 indicates that examinees' scaling scores differed from their true scores by an average of 0.89 standard deviation units and is thus very significant.

The separate calibration with SL linking and pair-wise calibration performed the best as demonstrated by small RMSDs for most of the six grades as compared to the three calibration and linking methods. Most RMSDs under these two methods were



about 0.28 to 0.29 units of ability parameters, except for grades 7 and 8 under pair-wise calibration. The separate calibration with MS linking yielded comparable RMSD as the separate with SL linking method and pair-wise calibration for grade 4 to grade 7. However, for grades 3 and 8, which are furthest away from the base grade 5, the RMSDs were somewhat higher than those under separate with SL linking and pair-wise calibration.

In addition, the SD of RMSDs for most grades under the separate calibration with MS linking is somewhat higher than that under the separate calibration with SL linking. For grades 3 and 8, the SDs with SL linking were 0.023 and 0.016, respectively. The SDs with MS linking were both 0.046. These higher SD values indicate the estimation of RMSD across replications was not stable when using the separate calibration with MS linking method for those grades that are furthest away from the base grade.

Under the unidimensional condition with varied grade distributions variances, the patterns of RMSDs were very similar to those conditions with fixed variances. The curves in Figures 4.3 and 4.4 are very similar, with only a small difference at the two ends. As an example, for the lower end in grades 3 and 4, the RMSDs under the varied variances condition were generally smaller than under the fixed variance condition for all five methods. For the base grade 5, there was almost no change in the RMSD values with any method. However, for grades 7 and 8, the RMSDs yielded by separate calibration with both SL and MS linking and pair-wise calibration increased slightly. The increased values ranged from about 0.01 to 0.03. In contrast, the RMSDs yielded by semi-concurrent calibration decreased slightly. There was little change in the RMSDs under the concurrent calibration for grade 7 and grade 8.

In summary, for the base grade 5, all five methods provided less biased results than any other grades under unidimensional conditions with both fixed and varied variances. Among the five methods, the concurrent calibration and semi-concurrent calibration yielded slightly less biased and more consistent estimates than the other three methods. For grades 4 and 6, the five calibration and linking methods also provided comparative estimates with the separate calibration with SL linking and pair-wise calibration performed slightly better than the other three.

For grades 3, 7, and 8, the five methods behaved quite differently, with the separate calibration with SL linking having the best performance. The pair-wise calibration and the separate calibration with MS linking behaved very similarly and were slightly worse than separate calibration with SL linking method. Concurrent calibration performed worse than the above three. Semi-concurrent calibration performed the worst among five methods based on the RMSDs. For grades 6, 7, and 8, the RMSDs were much higher for semi-concurrent calibration when compared to concurrent calibration. The reason is that the ability transformation method was used to put the two halves on one common scale rather than SL or Mean/sigma methods. This indicates the ability transformation method is not good in vertical scaling context.

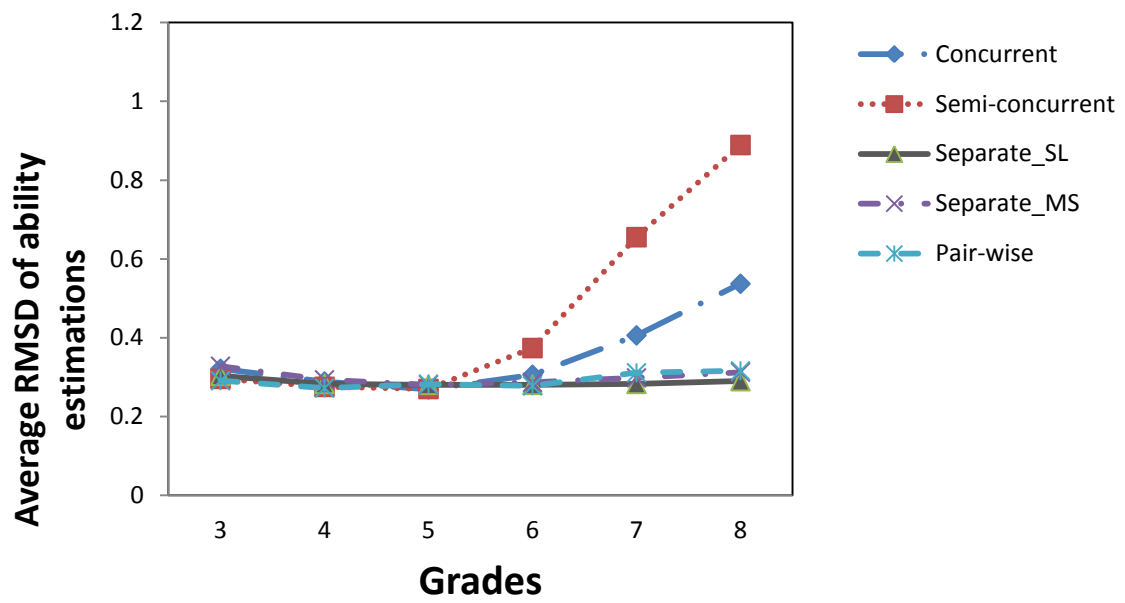
In addition, the SD of each method increased as grades were further from the base grade except for the separate calibration with SL linking. This indicates that the ability estimations were getting less stable when the grades were further from the base grade under the other four methods. Furthermore, the variances of each grade ability distribution only had a small impact on the estimations of latent abilities.

**Table 4.3** Average RMSDs of  $\hat{\theta}$  under unidimensional condition with fixed variance

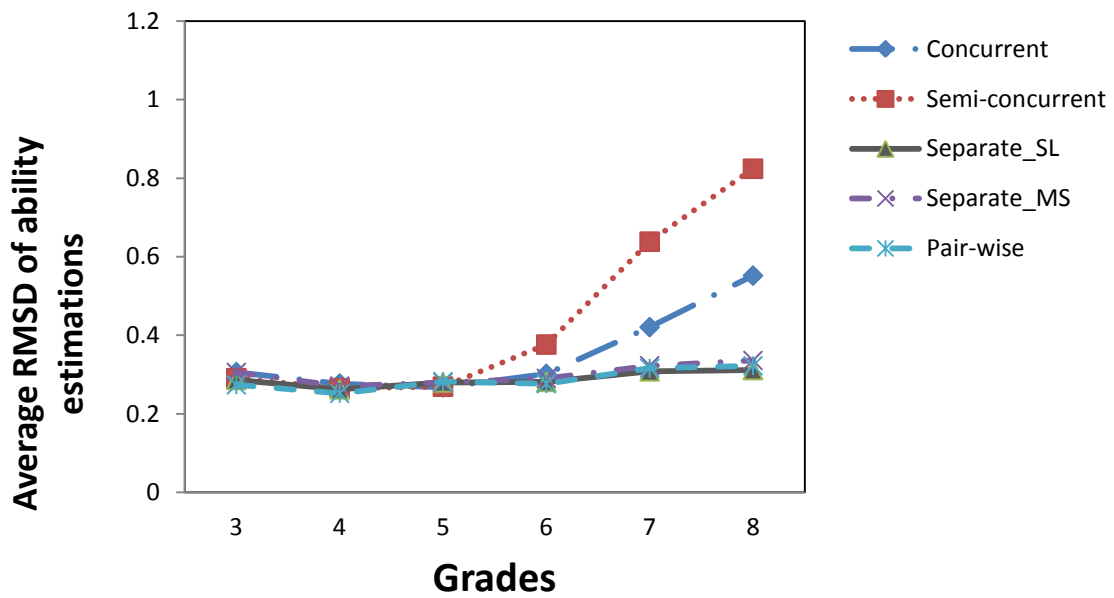
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.321(0.018)	0.296(0.012)	0.303(0.023)	0.328(0.046)	0.291(0.014)
Grade 4	0.289(0.014)	0.275(0.010)	0.283(0.017)	0.293(0.026)	0.272(0.013)
Grade 5	0.269(0.008)	0.270(0.008)	0.280(0.013)	0.280(0.013)	0.282(0.011)
Grade 6	0.306(0.016)	0.374(0.051)	0.280(0.014)	0.287(0.025)	0.278(0.013)
Grade 7	0.406(0.027)	0.655(0.126)	0.283(0.015)	0.298(0.036)	0.311(0.034)
Grade 8	0.537(0.033)	0.889(0.152)	0.290(0.016)	0.312(0.046)	0.316(0.035)

**Table 4.4** Average RMSDs of  $\hat{\theta}$  under unidimensional condition with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.306(0.012)	0.291(0.011)	0.286(0.021)	0.305(0.040)	0.274(0.014)
Grade 4	0.277(0.013)	0.265(0.009)	0.262(0.014)	0.271(0.021)	0.252(0.010)
Grade 5	0.267(0.008)	0.269(0.008)	0.280(0.013)	0.280(0.013)	0.282(0.009)
Grade 6	0.302(0.016)	0.376(0.056)	0.281(0.014)	0.290(0.025)	0.277(0.013)
Grade 7	0.421(0.027)	0.638(0.136)	0.308(0.017)	0.322(0.035)	0.316(0.033)
Grade 8	0.552(0.035)	0.824(0.133)	0.311(0.018)	0.336(0.050)	0.322(0.037)



**Figure 4.3** Average RMSDs of  $\hat{\theta}$  under unidimensional condition with fixed variance



**Figure 4.4** Average RMSDs of  $\hat{\theta}$  under unidimensional condition with varied variances

### **4.1.3 Correlation between estimated and “true” abilities**

The correlation between ability estimates and “true” abilities provides some information about the rank of examinees’ ability estimates. A higher correlation indicates that there is less of an impact on the rank of students’ latent ability regardless of the variances of the ability distributions. The correlations for the unidimensional condition with fixed and varied variances are reported in Tables 4.5 and 4.6.

Under the unidimensional condition with fixed variance, the correlations between ability estimates and “true” abilities were all generally high, with most correlations in the range of 0.960 to 0.965 (Table 4.5) and SD values about 0.002 and 0.003.

The concurrent calibration and separate calibration with SL linking and MS linking methods had the highest and most consistent correlations across the six grades. The correlations among the three methods were almost identical from grades 3 to 8. The correlations under pair-wise calibration were slightly lower than the correlations under the above three methods for grades 5 to 8. Most differences were about 0.003 to 0.004. The correlations under the semi-concurrent calibration for grades 3 to 5 were very close to the correlations under the other four calibration and linking methods. However, for grades 6 to 8, the correlations under the semi-concurrent calibration were much lower than the correlations under the other calibration methods, but the differences were only about 0.01 to 0.02 units.

Under the unidimensional condition with varied variances, the correlation patterns changed only slightly for all five calibration methods when compared to the correlations under the unidimensional condition with fixed variance. For base grade 5 and grade 6, the average correlations were almost the same under the two data structure

conditions for each calibration and linking methods. For grades 3 and 4, the correlations under the unidimensional condition with varied variances were slightly lower than the correlations under the fixed variance condition. Correlation values only dropped about 0.003 or 0.004. For grades 7 and 8, the correlations increased slightly with about 0.001 to 0.002 units.

In summary, under unidimensional conditions with either fixed or varied variances for ability distributions, the correlations were generally high except for a few under the semi-concurrent calibration. The separate calibration with SL linking and MS linking methods and concurrent calibration yielded the highest correlations for almost all six grades under both data structures. The pair-wise produced slightly lower correlations for grade 5 and higher. The correlations yielded by the semi-concurrent calibration were lower than the correlations produced by the other four methods for grades 6 to 8. However, for grades 3 to 5, the correlations yielded by semi-concurrent calibration were almost the same as the correlations from the four methods under the corresponding conditions. The bias and RMSD results indicate some spread between true ability and ability estimates, despite essentially equivalent rank order.

**Table 4.5** Average correlation of  $\theta$  and  $\hat{\theta}$  under unidimensional condition with fixed variance

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.959(0.003)	0.959(0.003)	0.959(0.003)	0.959(0.003)	0.959(0.003)
Grade 4	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.964(0.002)
Grade 5	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.960(0.003)
Grade 6	0.965(0.002)	0.946(0.007)	0.965(0.002)	0.965(0.002)	0.961(0.003)
Grade 7	0.965(0.002)	0.946(0.005)	0.964(0.002)	0.964(0.002)	0.961(0.003)
Grade 8	0.962(0.002)	0.941(0.006)	0.963(0.002)	0.963(0.002)	0.960(0.003)

**Table 4.6** Average correlation of  $\theta$  and  $\hat{\theta}$  under unidimensional condition with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.955(0.003)	0.955(0.003)	0.955(0.003)	0.955(0.003)	0.955(0.003)
Grade 4	0.962(0.002)	0.962(0.002)	0.962(0.002)	0.962(0.002)	0.962(0.002)
Grade 5	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.964(0.002)	0.960(0.003)
Grade 6	0.965(0.002)	0.946(0.007)	0.965(0.002)	0.965(0.002)	0.962(0.003)
Grade 7	0.966(0.002)	0.946(0.007)	0.966(0.002)	0.966(0.002)	0.963(0.002)
Grade 8	0.964(0.002)	0.943(0.006)	0.965(0.002)	0.965(0.002)	0.962(0.003)

#### 4.1.4 Summary

First, the different variance structures of ability distributions only had a small effect on the estimation of latent abilities. The bias patterns, RMSDs patterns, and correlation patterns across grades were very similar to each other with either fixed variance of 1 or varied variances for the six ability distributions.

Second, under unidimensional condition, the performance of the five calibration and linking methods were different. The average biases were very small across all grades when separate calibrations with SL and MS linking methods and semi-concurrent calibrations were used. With the concurrent calibration and semi-concurrent calibration, the average biases were also small for grades 3 to 6, but they were much larger for grades 7 and 8.

The average correlations were all similar and high. The average RMSDs were also generally small under most calibration and linking methods except for the concurrent calibration and semi-concurrent calibration at certain grades. For the base grade 5, all five methods provided the least biased results than any other grade. All RMSDs were about or smaller than 0.28 standard units and close to each other. For grades 3, 4, and 6, the average RMSDs were still small. The five calibration and linking methods provided comparable estimates, with the separate calibration with SL linking and pair-wise calibration performing slightly better than the other three methods. For grades 7 and 8, the five methods behaved very differently. The separate calibration with SL linking performed the best, followed by the pair-wise calibration and the separate calibration with MS linking. The concurrent calibration performed worse than the above three while the semi-concurrent calibration performed the worst. On the other hand, as



we can see in figures 4.1 and 4.2, the semi-concurrent calibration had smaller average bias than the concurrent calibration.

One reason for the different behavior in the average bias and the RMSD for the concurrent and the semi-concurrent calibrations is that average bias and RMSDs measure different properties about the bias, *i.e.*, the average bias provides only the direction of overall abilities, either overestimated or underestimated, while the RMSDs measure the magnitude of deviation from the average bias. For example, for grade 8 students, the abilities were overall highly underestimated under the concurrent calibration (negative direction), but the change in the variance of the ability distribution was relatively small. In contrast, under the semi-concurrent calibration, the variance of ability distribution was expanded while the average bias stayed relatively unchanged. So, even though the abilities were underestimated for students at lower level and overestimated for students at higher level, the overall average bias could be small since the negative and positive biases cancel each other in calculating the average bias. However, the RMSDs could be large because it is calculated based on the squared bias, which ignores the direction of the bias.

In general, under the unidimensional condition, the separate calibration with SL linking provided the least biased ability estimates across all six grades. The pair-wise calibration and separate calibration with MS linking resulted in comparable ability estimates as the separate calibration with SL linking. This finding is consistent with previous finding that the SL linking performed marginally superior in estimation accuracy compared to MS linking (Baker & Al-Karni, 1991; Stocking & Lord, 1983). The concurrent calibration performed also well for the middle grades, *i.e.*, the base grade 5 and the adjacent grades. However, for the grades furthest away from the base grades

such as grades 7 and 8, the concurrent calibration yielded more biased results. Overall, the abilities for the higher grades 7 and 8 were underestimated. One possible reason for this is that the Bayesian estimation method, MAP, was used for estimating latent abilities. This method pulled the abilities at very end toward the mean. The performance of the semi-concurrent calibration was very different across grades. For the lower level grades, the performance of the semi-concurrent calibration was very good. However, for the higher level grades, semi-concurrent calibration yielded very biased results, especially for grades 7 and 8.

## **4.2 COMPARISONS OF SCALING METHODS UNDER MULTIDIMENSIONAL CONDITON WITH $R = .3$ BETWEEN TWO LATENT ABILITIES**

### **4.2.1 Bias of ability estimates**

As mentioned in Chapter 3, the unidimensional condition is used for reference and comparison with the multidimensional conditions. The multidimensional conditions more likely reflect the real test situation in practice. There were two assumed multidimensional conditions in this study. One condition reflected a relatively low relationship between the dominant ability and the secondary ability ( $r=0.3$ ). In the other condition, the secondary ability was moderately related to the dominant ability ( $r=0.6$ ).

The bias results under the multidimensional condition ( $r=0.3$ ) with two variance structures for the six ability distributions are reported in Tables/Figures 4.7 and 4.8. For the multidimensional condition ( $r=0.3$ ) with fixed variance of 1 across all ability distributions, some bias values were positive and some of them were negative (Table 4.7 and Figure 4.7). This indicates that some of the average estimated abilities were larger than the “true” abilities and some were smaller than the “true” abilities. The degree of deviation was different among various calibration and scaling methods.

Compared to the results under the unidimensional condition for most calibration and scaling methods, the average biases under the multidimensional data condition were different. Specifically, the average biases changed when the grades were further away from the base grade 5.

For the separate calibrations with both SL linking and MS linking and pair-wise calibration, the average biases became negative for the lower grade 3, with the absolute values being small. However, the average biases increased for grades 6, 7, and 8 with most of them being larger than 0.1 units. The semi-concurrent calibration and concurrent calibrations had similar bias patterns as under the unidimensional condition where the average biases were still large for grades 7 and 8. Among the five calibration and linking methods, the separate calibration with SL linking and MS linking and the pair-wise calibration still provided the smallest average biases for most grades, with SL linking being slightly better than the other two.

For the multidimensional condition ( $r=0.3$ ) with varied variances for ability distributions, the bias patterns under all five calibration methods were similar to those for the multidimensional condition ( $r=0.3$ ) with fixed variance (see Figures 4.5 and 4.6). For the separate calibration with SL linking and MS linking, the average biases were slightly larger than those under the multidimensional condition ( $r=0.3$ ) with fixed variance for grade 8. The other bias differences were about 0.01 to 0.02 units. For the concurrent calibration and semi-concurrent calibration, the bias differences were also small except for grade 3, most bias differences were about 0.01 units. These differences were small compared to the average biases for each grade which were produced by each calibration and scaling methods.

In general, the average biases under the multidimensional condition ( $r=0.3$ ) were different from the biases under the unidimensional condition. Under the unidimensional condition, most average biases were very small with most values being very close to 0 except for bias results produced by semi-concurrent and concurrent calibrations in grades 7 and 8. Under the multidimensional condition ( $r=0.3$ ) with both fixed variance

and varied variance conditions, the absolute average biases under the five calibration and linking methods were overall larger for most grades, especially for grades 7 and 8.

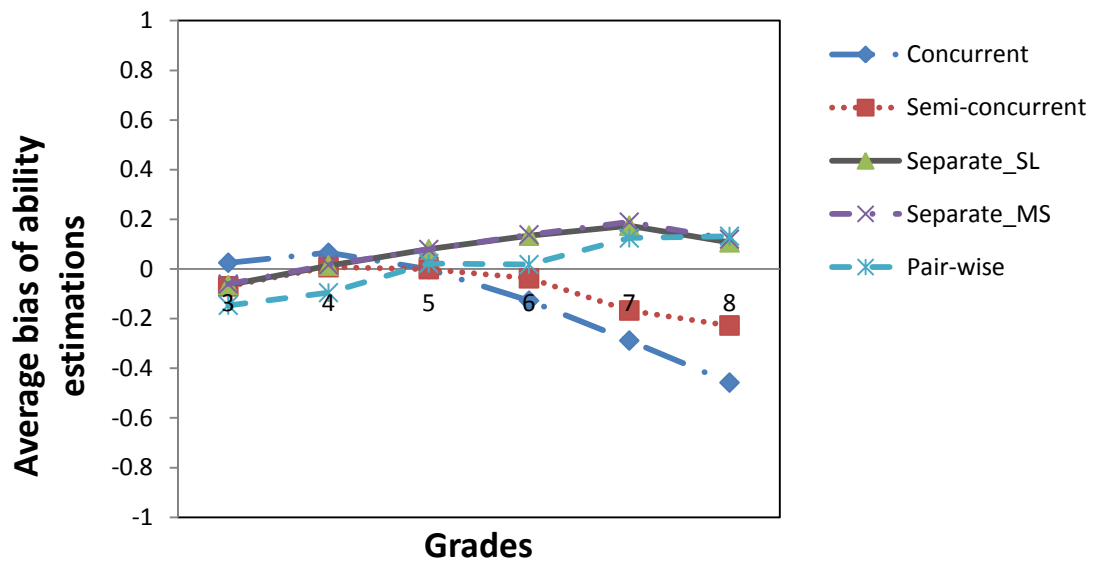
For the base grade 5 and the adjacent grade 4, the average biases were generally much smaller than those for the other grades. For the grades furthest away from the base grade 5, such as grades 7 and 8, all five calibration and linking methods provided larger average biases. The abilities were overall underestimated under concurrent calibration and semi-concurrent calibration. However, the abilities were somewhat overestimated under separate calibration with SL linking and MS linking and pair-wise calibration.

**Table 4.7** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance

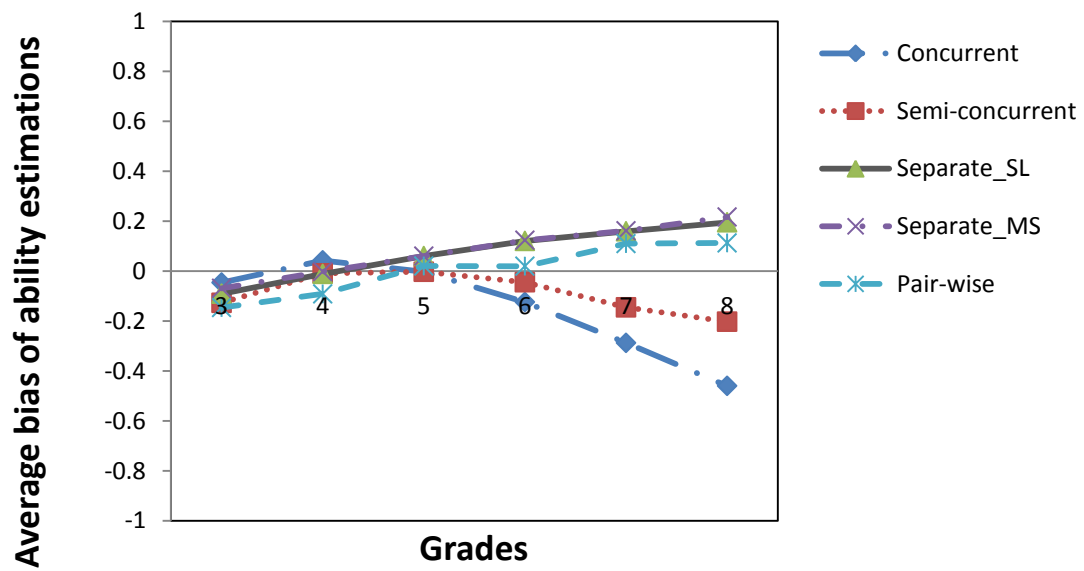
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.025(0.047)	-0.070(0.037)	-0.065(0.064)	-0.061(0.120)	-0.147(0.060)
Grade 4	0.065(0.038)	0.007(0.032)	0.014(0.050)	0.016(0.065)	-0.095(0.053)
Grade 5	-0.001(0.024)	-0.001(0.024)	0.080(0.041)	0.079(0.046)	0.022(0.025)
Grade 6	-0.128(0.039)	-0.037(0.089)	0.134(0.044)	0.138(0.065)	0.018(0.028)
Grade 7	-0.288(0.044)	-0.167(0.169)	0.174(0.048)	0.189(0.093)	0.125(0.069)
Grade 8	-0.458(0.047)	-0.228(0.199)	0.107(0.147)	0.120(0.182)	0.132(0.077)

**Table 4.8** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	-0.046(0.047)	-0.127(0.039)	-0.091(0.062)	-0.069(0.136)	-0.146(0.055)
Grade 4	0.043(0.035)	-0.009(0.030)	-0.011(0.039)	-0.002(0.065)	-0.092(0.037)
Grade 5	-0.003(0.021)	-0.003(0.021)	0.060(0.028)	0.060(0.028)	0.020(0.024)
Grade 6	-0.124(0.035)	-0.045(0.091)	0.121(0.038)	0.124(0.053)	0.020(0.029)
Grade 7	-0.288(0.039)	-0.145(0.140)	0.159(0.044)	0.162(0.079)	0.110(0.078)
Grade 8	-0.460(0.043)	-0.203(0.190)	0.195(0.051)	0.217(0.120)	0.113(0.081)



**Figure 4.5** Average bias of  $\hat{\theta}$  at multidimensional condition ( $r=0.3$ ) with fixed variance



**Figure 4.6** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with varied variances

#### 4.2.2 RMSD of ability estimates

The RMSD for the multidimensional condition ( $r=0.3$ ) with fixed and varied variances for ability distributions are reported in Table 4.9 and 4.10, respectively. Similar to section 4.1.2, in each table, the average RMSD of ability estimates were compared for the five calibration and linking methods from grades 3 to 8. The value in each cell is the average RMSD of each grade by each method over 200 replications. The value in the parenthesis is the standard deviation (SD) of the RMSD estimates.

Under the multidimensional condition ( $r=0.3$ ) with fixed variance, the average RMSDs and SD of RMSDs at each grade under different calibration and scaling methods were generally greater than those values provided by corresponding methods under the unidimensional condition.

First, the average RMSD values increased significantly for almost all grades under this multidimensional condition. For the base grade 5, almost all average RMSDs were about 0.27 to 0.28 standard units for the base grade 5 for all five calibration and scaling methods under unidimensional condition. Under the multidimensional condition, the average RMSDs were about 0.34 to 0.35 for the base grade. So, the average RMSDs from each calibration and linking methods increased by about 0.07 units compared to under the unidimensional condition. For the two adjacent grades, grades 4 and 6, the RMSDs generally increased by about 0.05 to 0.09 standard units for most calibration and linking methods compared to corresponding grades under the unidimensional condition. For grades 3, 7, and 8, the RMSDs changed differently with different calibration and linking methods under this multidimensional condition. For concurrent calibration and semi-concurrent calibration, the RMSDs did not change much for these grades. Most



differences were about 0.01 units. For separate calibration with SL linking and MS linking and pair-wise calibration, the RMSDs increased much more, some even up to 0.1 units when compared to unidimensional condition.

Second, the standard deviation (SD) of RMSDs at each cell were larger except for the base grade 5, for which the SDs of RMSDs were the similar under multidimensional and unidiemnsional conditions. However, the SEs increased more quickly for the grades further away from the base grade such as grades 7 and 8, especially for the separate calibration with MS linking. For example, the SD value was 0.046 for grade 8 under the unidimensional condition with fixed variance of 1 using separate calibration with MS linking, and it increased to 0.080 under the multidimensional condition. It was about two times the SD for this grade under the unidimensional condition and was almost six times the SD at the base grade 5 (SD=0.013) using the same method. The relatively high SEs indicate that the separate calibration with MS linking method was fairly sensitive to this multidimensional data structure.

Third, among the five calibration and linking methods, the average RMSDs under the separate calibration and SL linking were still the smallest across most grades. In addition, the line for the separate calibration and SL linking was fairly flat (Figure 4.7). It indicates that the estimations under separate calibration with SL linking were still relatively consistent across grades even the overall RMSDs increased. The separate calibration with MS linking and the pair-wise calibration yielded slightly larger RMSDs than the separate calibration with SL linking, especially for the grades furthest away from the base grade 5. The concurrent and semi-concurrent calibration had similar and comparative RMSDs as the above three methods for the lower grades 3 to 6. However,

they had much larger RMSDs for higher grades such as grades 7 and 8. The semi-concurrent calibration had the largest RMSDs for the two grades even the average biases yielded by semi-concurrent calibration were smaller than concurrent calibration did. For example, for grade 8, the average bias was -0.280 units, while the average value for RMSDs was very close to 0.9. On average, the ability estimate was one standard deviation from the “true” ability, indicting it is a biased ability estimate. This can also be verified from Figure 4.9 because the right tail of the semi-concurrent method was much higher than the tails of the other four methods.

One possible reason for this is that some abilities were underestimated while some abilities were overestimated under semi-concurrent calibration. Because the average bias is the average of 2000 biases of ability estimates and across 200 replications (note: it was not absolute bias, it could be positive or negative). The average bias just provided information about if, and by how much, the overall abilities were overestimated or underestimated. Only the average RMSDs provides information about the magnitude of bias, since it is the sum of squared bias and averaged by 2000 samples and across 200 replications, then taking square root of it. For example, when the variance of ability distribution of grade 8 was expanded under semi-concurrent calibration, the abilities were underestimated for students at lower grade levels while the abilities were overestimated for students at higher grade levels. Overall, the average bias is small since some biases are negative while the others are positive. The RMSDs were large only if the variances increased significantly. Similarly, under the concurrent calibration, the ability distribution of grade 8 was shifted to the left because, under the MAP ability estimation method, a Bayesian method, ability estimates are biased and toward the mean when all 6 grades calibrated together. Overall, the abilities were

underestimated for almost each examinee. The average biases are larger under semi-concurrent calibration because almost all of them are negative. The RMSDs under concurrent calibration are smaller than under semi-concurrent calibration when the distribution did not shift much.

In general, the order of in magnitude of RMSDs produced by the five calibration and linking methods from the smallest to highest are: separate calibration with SL linking, pair-wise calibration, separate calibration with MS linking, concurrent calibration, and semi-concurrent calibration.

Under the multidimensional condition ( $r=0.3$ ) with varied variances, most of the average RMSDs for middle grades such as grades 4 to 6 produced by the five calibration and linking methods were similar to the RMSDs for the multidimensional condition ( $r=0.3$ ) with fixed variance (Table 4.10 and Figure 4.8). These differences were in the range 0.01 to 0.02 standard units.

However, for grades 7 and 8, the average RMSDs under the multidimensional condition ( $r=0.3$ ) with varied distribution variances were even larger than the corresponding average RMSDs under the multidimensional condition ( $r=0.3$ ) with fixed distribution variance for most calibration and linking methods, especially for the separate calibration with linking methods. For example, for grade 8 under the multidimensional ( $r=0.3$ ) with fixed distribution variance condition, the RMSDs were 0.353 and 0.386 standard units for separate calibration with SL linking and MS linking, respectively. Under the multidimensional ( $r=0.3$ ) with varied distribution variances condition, the corresponding RMSDs were 0.382 and 0.43 standard units. The average RMSDs increased 0.03 and 0.04 for SL linking and MS linking, respectively. That is, the separate calibrations with linking methods were influenced slightly more by the ability

distributions with varied variances for different grades than the other methods did when the grades were furthest away from the base grade.

In summary, the average RMSDs under this multidimensional condition ( $r=0.3$ ) were generally larger than the RMSDs for corresponding grades and methods under the unidimensional condition. For the base grade 5, the average RMSDs increased by similar amount from those under the unidimensional condition for all five calibration and linking methods. The resulted RMSDs were very close to each other for the five methods. For grades 3, 4 and 6, the amount of increase in the average RMSDs is slightly different for the five methods, but the results are still comparable. For grades 7 and 8, the RMSDs were very different for the five calibration and linking methods. The RMSDs under the separate calibration with SL linking method increased consistently as the middle grades did. Those under the separate calibration with MS linking method and pair-wise calibration were slightly larger than separate calibration with SL linking. It indicates that the SL linking (or TCC method) is marginally superior in estimation accuracy compared to MS linking. There was a small increase on the RMSDs with the concurrent calibration and semi-concurrent calibration. The average RMSDs were still much larger than the separate calibration with SL linking and MS linking and pair-wise calibration. In general, these larger values of RMSDs indicate that the multidimensional structure did have an effect on the ability estimates.

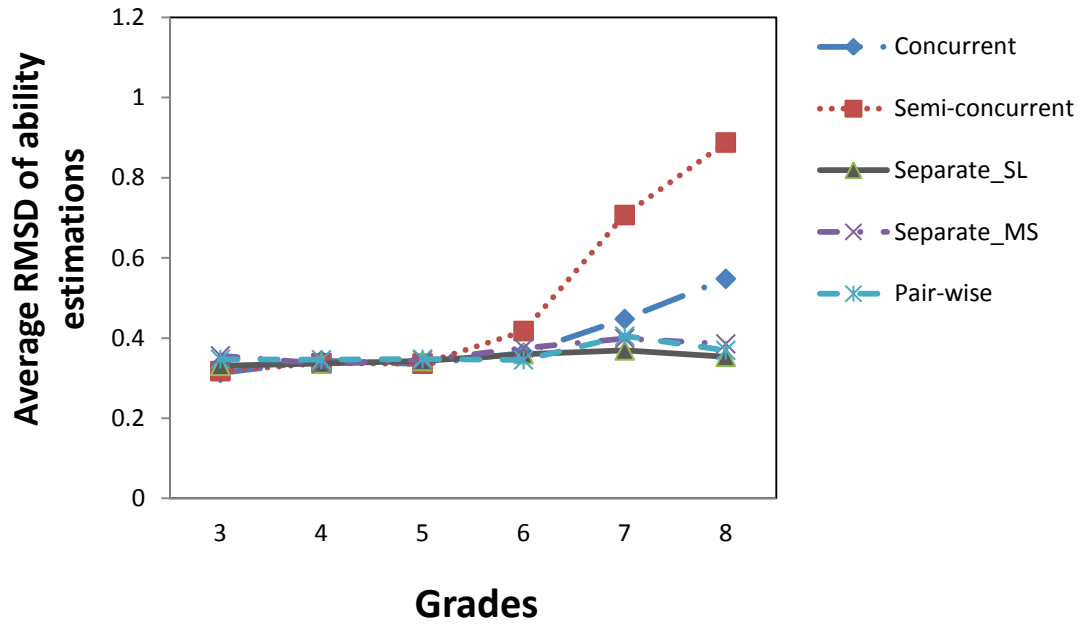
The average RMSD values under the multidimensional condition ( $r=0.3$ ) were generally similar with both fixed and varied variances for the ability distribution, indicating the differences in ability variances had a small effect on the ability estimates. The largest effect was for the separate calibration with MS linking at grade 8 which is the furthest grade from the base grade.

**Table 4.9** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance

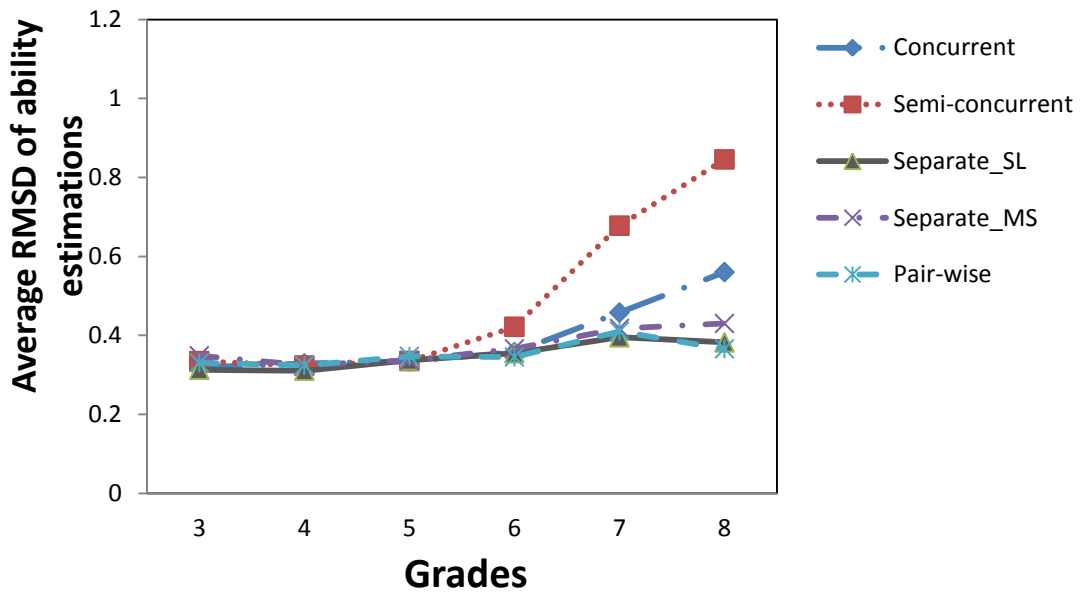
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.313(0.012)	0.318(0.015)	0.331(0.015)	0.356(0.057)	0.347(0.030)
Grade 4	0.344(0.012)	0.337(0.010)	0.337(0.010)	0.346(0.022)	0.351(0.025)
Grade 5	0.335(0.009)	0.336(0.009)	0.343(0.013)	0.345(0.018)	0.348(0.010)
Grade 6	0.363(0.018)	0.418(0.039)	0.360(0.018)	0.375(0.039)	0.345(0.013)
Grade 7	0.448(0.029)	0.707(0.123)	0.369(0.023)	0.399(0.063)	0.405(0.046)
Grade 8	0.548(0.040)	0.888(0.156)	0.353(0.030)	0.386(0.080)	0.369(0.050)

**Table 4.10** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.319(0.019)	0.334(0.022)	0.313(0.019)	0.348(0.062)	0.331(0.024)
Grade 4	0.329(0.012)	0.323(0.011)	0.310(0.009)	0.325(0.025)	0.324(0.014)
Grade 5	0.334(0.009)	0.335(0.009)	0.337(0.010)	0.337(0.010)	0.347(0.010)
Grade 6	0.358(0.015)	0.422(0.044)	0.355(0.016)	0.366(0.031)	0.345(0.016)
Grade 7	0.458(0.027)	0.678(0.091)	0.395(0.020)	0.418(0.052)	0.410(0.036)
Grade 8	0.560(0.036)	0.846(0.134)	0.382(0.026)	0.430(0.126)	0.366(0.039)



**Figure 4.7** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance



**Figure 4.8** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with varied variances

### 4.2.3 Correlation between estimated and “true” abilities

The correlations of estimated and “true” ability distributions under the multidimensional condition with fixed variance and varied variances are reported in Tables 4.11 and 4.12, respectively.

Under the multidimensional condition ( $r=0.3$ ) with fixed variance, the correlations between ability estimates and “true” abilities were similar across the five methods, but slightly lower than the corresponding correlations under the unidimensional condition. More specifically, under the unidimensional condition, most correlations were about 0.964, while under this multidimensional condition ( $r=0.3$ ), most correlations were about 0.944. The slightly decreased correlations further verified the multidimensionality on common items had some effect on ability estimation.

The SEs of the correlations under the multidimensional condition were also very small and slightly higher than those under the unidimensional condition, with most values falling between 0.003 and 0.004. This indicates the estimates of correlations were still fairly stable under multidimensional data structure even the overall correlations dropped.

Under multidimensional ( $r=0.3$ ) with varied variances condition, the correlation matrix changed slightly when compared to the fixed variance condition. Most correlations were very similar for the base grade 5 and grade 6. For grades 3 and 4, correlations decreased by a very small amount, *i.e.*, about 0.004 or 0.005 units under each calibration and scaling methods, while for grades 7 and 8, the correlations increased by 0.002 to 0.004 units under each method. This small change may be due to the smaller variances for grades 3 and 4 ability distributions and larger variances for

grades 7 and 8 ability distributions. In general, these differences were much smaller than the differences due to the difference in dimensionality.



**Table 4.11** Average correlation of  $\theta$  and  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.953(0.004)	0.953(0.004)	0.950(0.004)	0.950(0.006)	0.953(0.003)
Grade 4	0.943 (0.003)	0.944(0.003)	0.944(0.003)	0.944(0.003)	0.943(0.003)
Grade 5	0.944(0.003)	0.944(0.003)	0.944(0.003)	0.944(0.003)	0.939(0.003)
Grade 6	0.944(0.003)	0.927(0.007)	0.944(0.003)	0.944(0.003)	0.940(0.003)
Grade 7	0.945(0.003)	0.923(0.009)	0.948(0.006)	0.948(0.006)	0.940(0.003)
Grade 8	0.956(0.002)	0.933(0.007)	0.955(0.003)	0.955(0.003)	0.953(0.003)

**Table 4.12** Average correlation of  $\theta$  and  $\hat{\theta}$  under multidimensional condition( $r=0.3$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.947(0.004)	0.947(0.004)	0.948(0.004)	0.948(0.006)	0.948(0.003)
Grade 4	0.940 (0.003)	0.940(0.003)	0.940(0.003)	0.940(0.003)	0.940(0.003)
Grade 5	0.944(0.003)	0.944(0.003)	0.944(0.003)	0.944(0.003)	0.939(0.004)
Grade 6	0.944(0.003)	0.926(0.007)	0.944(0.003)	0.944(0.003)	0.940(0.004)
Grade 7	0.947(0.003)	0.926(0.006)	0.947(0.003)	0.947(0.003)	0.942(0.003)
Grade 8	0.958(0.002)	0.936(0.006)	0.959(0.002)	0.959(0.002)	0.956(0.003)

#### 4.2.4 Summary

First, similar to the unidimensional condition, the different variance structures of ability distributions only had a small effect on the estimates of latent abilities. The bias patterns, RMSDs patterns, and correlation patterns across grades were generally similar to each other under the two variance structures for the six ability distributions. Compared to the fixed variance structures, the biases of ability estimates were slightly larger for the higher grades and slightly lower for the lower grades for the varied variance structure condition.

Second, compared to unidimensional condition, the patterns of average biases and RMSDs under the multidimensional condition ( $r=0.3$ ) were generally greater, *i.e.* most biases and RMSDs were larger for most calibrations and scaling methods except for concurrent calibration at certain grades on the left tail. Similar to the unidimensional condition, all five methods produced the least biased results for the base grade 5, with all RMSDs being close to each other. For grades 3, 4, and 6, the five calibration and linking methods provided comparable estimates, with the separate calibration with SL linking and the pair-wise calibration performing slightly better than the other three methods. Finally, for grades 7 and 8, most RMSDs were significantly higher than the other grades except for the separate calibration with SL linking method, with semi-concurrent calibration yielding the largest RMSDs.

Although correlations between “true” abilities and ability estimates were similar and high, the average correlations at each grade were slightly lower than those under the corresponding unidimensional condition. The slightly decrease on correlations also indicated that the introduction of multidimensionality affected ability estimations.

### 4.3 COMPARISONS OF SCALING METHODS UNDER MULTIDIMENSIONAL CONDITON WITH $R = .6$ BETWEEN TWO LATENT ABILITIES

#### 4.3.1 Bias of ability estimates

The bias results for the multidimensional condition with a moderate relationship between the dominant ability and the secondary ability ( $r=0.6$ ) are reported in this section. Tables 4.13 and 4.14 provided the biases results for the multidimensional condition ( $r=0.6$ ) with fixed variance of 1 and varied variances for ability distributions, respectively. Similar to sections 4.1.1 and 4.2.1, in each table, the biases of ability estimates,  $\hat{\theta}_i - \theta_i$ , were compared for the five calibration and linking methods from grades 3 to 8.

Under the multidimensional condition ( $r=0.6$ ) with fixed variance of 1 for ability distributions, the average biases for the grades 3 to 6 were still low, *i.e.* for these grades, most average biases were still under 0.1 units and close to 0.001. Similar to the multidimensional condition ( $r=0.3$ ), the averages biases increased for the grades 7 and 8 under separate calibration with two linking methods compared to the unidimensional condition. The average biases under separate calibrations were over 0.1 units. There was a very small change on average biases for the other three calibration and linking methods. The patterns of average bias under the two multidimensional conditions ( $r=0.3$  and  $r=0.6$ ) were similar, with some average biases being slightly smaller than those under the multidimensional condition ( $r=0.3$ ).

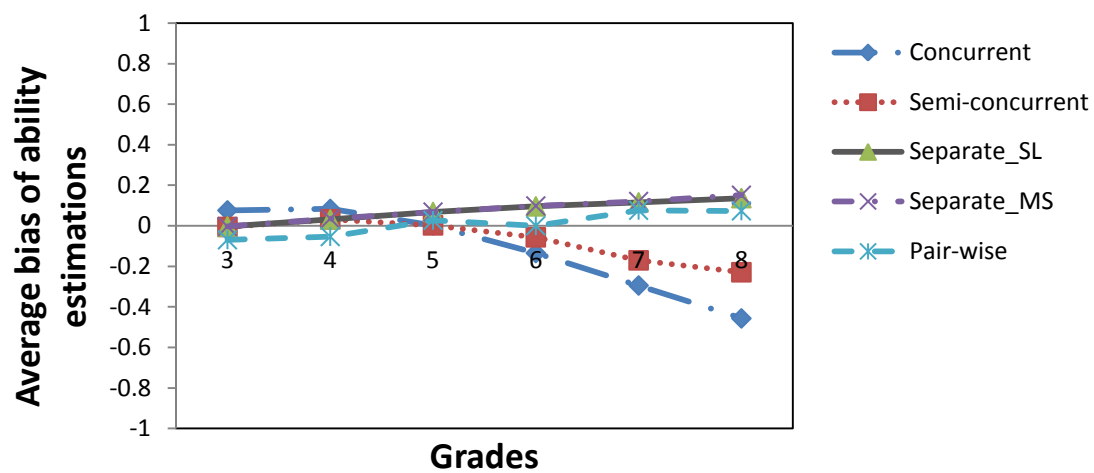
Similar to the results shown in section 4.1.1 and 4.2.1, the bias with varied variances for ability distributions were almost identical to those with fixed variance of 1 under the multidimensional condition ( $r=0.6$ ) . The variances of ability distributions had small effects on average biases.

**Table 4.13** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance

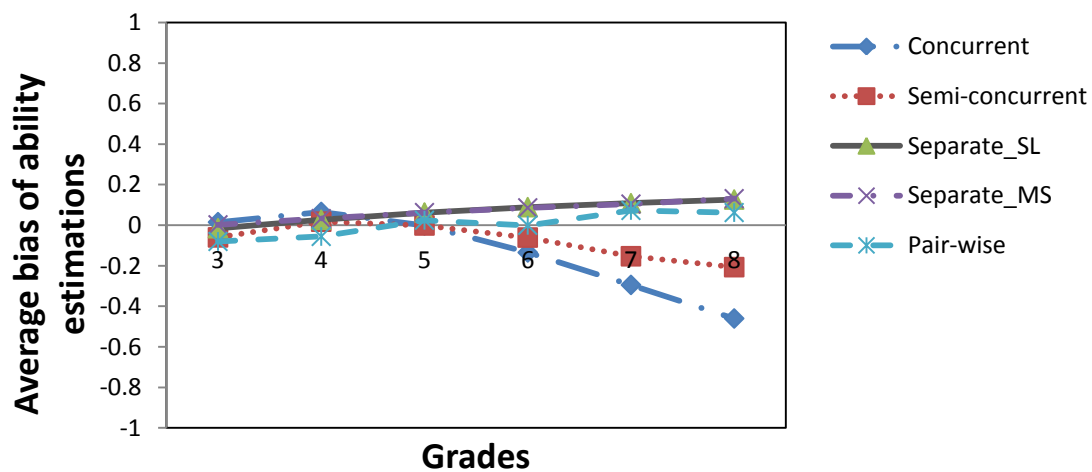
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.076(0.045)	-0.006(0.037)	-0.003(0.056)	-0.003(0.113)	-0.069(0.054)
Grade 4	0.083(0.013)	0.032(0.009)	0.032(0.011)	0.035(0.026)	-0.053(0.012)
Grade 5	0.002(0.023)	0.002(0.023)	0.068(0.031)	0.068(0.031)	0.027(0.025)
Grade 6	-0.134(0.035)	-0.057(0.089)	0.096(0.039)	0.098(0.050)	0.000(0.031)
Grade 7	-0.294(0.043)	-0.170(0.154)	0.116(0.046)	0.120(0.067)	0.076(0.083)
Grade 8	-0.457(0.044)	-0.229(0.187)	0.137(0.051)	0.151(0.097)	0.073(0.088)

**Table 4.14** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.015(0.042)	-0.060(0.036)	-0.015(0.054)	0.002(0.111)	-0.080(0.053)
Grade 4	0.065(0.032)	0.017(0.028)	0.027(0.042)	0.033(0.058)	-0.056(0.037)
Grade 5	-0.001(0.023)	-0.001(0.023)	0.062(0.030)	0.062(0.030)	0.024(0.024)
Grade 6	-0.135(0.031)	-0.061(0.092)	0.089(0.039)	0.086(0.046)	-0.001(0.030)
Grade 7	-0.294(0.039)	-0.153(0.131)	0.108(0.047)	0.104(0.063)	0.073(0.084)
Grade 8	-0.460(0.044)	-0.207(0.166)	0.127(0.054)	0.131(0.085)	0.062(0.087)



**Figure 4.9** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance



**Figure 4.10** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with varied variances

### 4.3.2 RMSD of ability estimates

The RMSD for the multidimensional condition ( $r=0.6$ ) with fixed variance 1 for all ability distributions and varied variances are reported in Tables 4.15 and 4.16, respectively. Similar to section 4.1.2 for the unidimensional condition and section 4.2.2 for the multidimensional condition with  $r=0.3$ , the RMSDs were very similar for each calibration and linking method with fixed variance of 1 and varied variances. Differences of the average RMSDs were less than 0.01 for most calibration and linking methods and grades. This once again indicates that the different variance structures had a relatively small effect on the ability estimates.

Under this multidimensional condition ( $r=0.6$ ), the average RMSDs were different from those under the unidimensional condition. Similar to the multidimensional condition with  $r=0.3$ , almost all RMSDs increased for the five calibration and linking methods. The average RMSD values under this multidimensional condition with  $r=0.6$  increased slower than those under the other multidimensional condition. For the base grade 5, the average RMSDs were about 0.27 to 0.28 for the five calibration and linking methods under unidimensional condition. Under the multidimensional with  $r=0.3$  condition, the average RMSDs were about 0.34 to 0.35, with an increase of about 0.07 units from the unidimensional condition. However, under the unidimensional condition with  $r=0.6$ , the average RMSDs were 0.31 to 0.32, indicating an increase of about 0.04 units when compared to the unidimensional condition. For the grades 4 and 6, the RMSDs increased by a similar amount as the grade 5 did for most calibration and linking methods except for semi-concurrent calibration. For grade 3, the RMSDs for each method increased by slightly smaller values. For grades 7 and 8, the RMSDs increased

differently for the five methods. The RMSDs under separate calibration with linking methods increased more rapidly than the concurrent calibration and semi-concurrent calibration, while the pair-wise calibration was in the middle. The separate calibration with MS linking method increased by the most amount.

However, among the five calibration and linking methods, semi-concurrent calibration still had the largest RMSDs. The separate calibration with SL linking had the smallest RMSDs. The separate calibration and the pair-wise calibration had slightly larger RMSDs than the separate calibration with SL linking. The concurrent calibration had very comparable RMSDs as the separate calibration with SL linking for grades 3 to 6. For grade 7 and 8, it provided more biased ability estimates.

The RMSDs patterns can be examined in Figures 4.11 and 4.12. The left tails of the five methods were flat and almost identical. However, the right tail of the semi-concurrent calibration was much higher than those for the other four methods. The concurrent calibration was in the middle. The separate calibration with SL linking was flat and the lowest across grades. The difference between the separate calibration with MS and SL linking become slightly larger than under the unidimensional condition.

There are two important implications from these RMSDs trends. First, the results provide evidence that the multidimensional data affected the ability estimates for the five calibrations with linking methods. Second, the results imply that the separate calibration with the MS linking method was affected more by multidimensional structure than the other methods. The separate calibration with SL linking was more robust to unidimensional violation than the other methods.

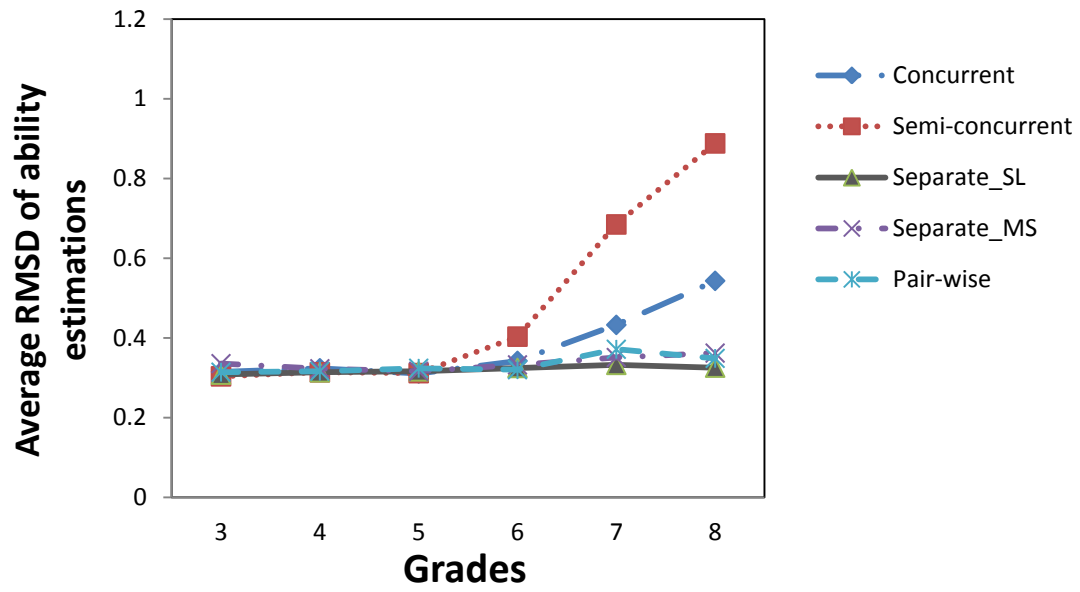


**Table 4.15** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance

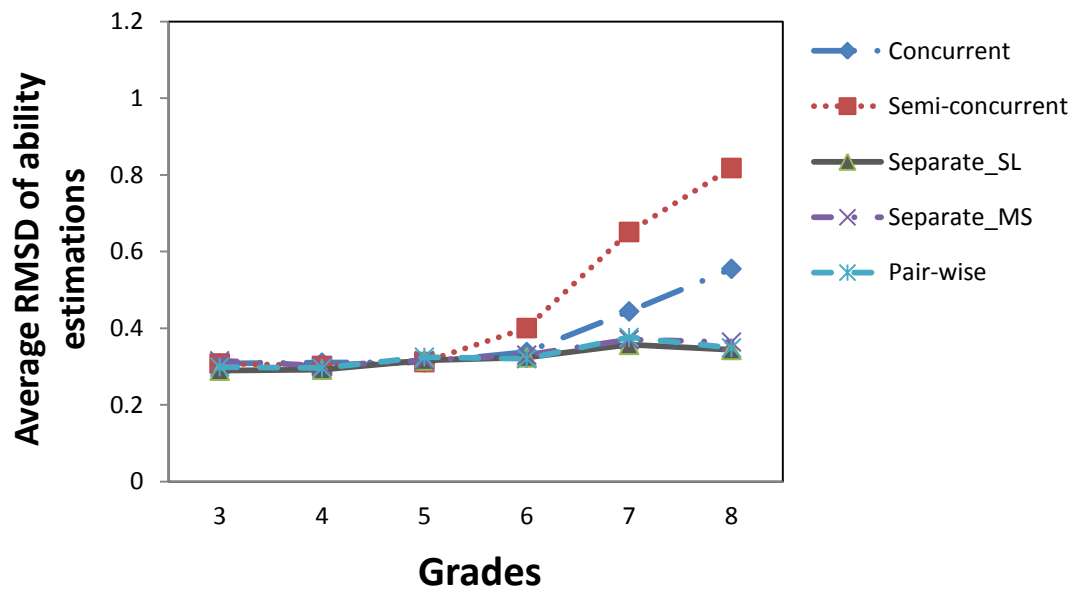
Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.315(0.013)	0.303(0.010)	0.308(0.014)	0.335(0.051)	0.314(0.017)
Grade 4	0.324(0.013)	0.314(0.009)	0.314(0.011)	0.323(0.026)	0.316(0.012)
Grade 5	0.311(0.008)	0.312(0.008)	0.316(0.010)	0.316(0.010)	0.324(0.010)
Grade 6	0.342(0.016)	0.403(0.043)	0.325(0.014)	0.332(0.024)	0.320(0.011)
Grade 7	0.433(0.029)	0.685(0.109)	0.332(0.017)	0.351(0.044)	0.371(0.043)
Grade 8	0.543(0.037)	0.888(0.147)	0.326(0.021)	0.362(0.077)	0.349(0.047)

**Table 4.16** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.308(0.012)	0.308(0.015)	0.290(0.013)	0.315(0.041)	0.299(0.017)
Grade 4	0.310(0.012)	0.302(0.010)	0.292(0.011)	0.302(0.023)	0.296(0.013)
Grade 5	0.311(0.008)	0.312(0.008)	0.316(0.010)	0.316(0.010)	0.325(0.012)
Grade 6	0.338(0.015)	0.401(0.043)	0.324(0.013)	0.330(0.026)	0.321(0.013)
Grade 7	0.444(0.027)	0.651(0.109)	0.357(0.018)	0.370(0.035)	0.376(0.035)
Grade 8	0.555(0.036)	0.818(0.144)	0.344(0.022)	0.364(0.047)	0.349(0.037)



**Figure 4.11** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance



**Figure 4.12** Average RMSD of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with varied variances

### 4.3.3 Correlation between estimated and “true” abilities

The correlations for the multidimensional condition ( $r=0.6$ ) with fixed variance 1 for all ability distributions and varied variances are reported in Tables 4.17 and 4.18, respectively. Similar to sections 4.1.3 and 4.2.3, the average correlations were very similar across of fixed and varied variance conditions for each grade. The SEs of the correlations under the multidimensional condition with  $r=0.6$  were also very small and similar to those under both the unidimensional and the multidimensional condition with  $r=0.3$ . Most values were about 0.002 and 0.003.

Similar to the multidimensional condition with  $r=0.3$ , most correlations were slightly lower than the corresponding values under the unidimensional conditions. More specifically, under the unidimensional condition, most correlations were about 0.965 with the exception of grade 3 and some correlations when using semi-concurrent calibration. Under the multidimensional condition with  $r=0.3$ , most correlations were about 0.944. Under multidimensional condition with  $r=0.6$ , most correlations were about 0.952, which was between the correlation values for the unidimensional condition and multidimensional condition with  $r=0.3$ .

In summary, as discussed earlier, the different ability variance structures had a very small effect on the correlations between ability estimates and “true ability”. Dimensionality had some effect on the correlations between ability estimates and “true” abilities. The average correlations under the multidimensional condition ( $r=0.6$ ) decreased slightly when compared to the unidimensional condition, but less difference was observed when comparing the multidimensional condition ( $r=0.3$ ) with the unidimensional condition.

**Table 4.17** Average correlation of  $\theta$  and  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.955(0.003)	0.955(0.003)	0.955(0.003)	0.955(0.003)	0.955(0.003)
Grade 4	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.952(0.003)
Grade 5	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.947(0.003)
Grade 6	0.952(0.002)	0.934(0.006)	0.952(0.002)	0.952(0.002)	0.948(0.003)
Grade 7	0.953(0.003)	0.933(0.006)	0.952(0.003)	0.952(0.003)	0.948(0.003)
Grade 8	0.958(0.002)	0.937(0.006)	0.959(0.002)	0.959(0.002)	0.956(0.002)

**Table 4.18** Average correlation of  $\theta$  and  $\hat{\theta}$  under multidimensional condition( $r=0.6$ ) with varied variances

Grade	Concurrent calibration(SD)	Semi-concurrent calibration (SD)	Separate with SL linking (SD)	Separate with MS linking (SD)	Pair-wise calibration(SD)
Grade 3	0.950(0.004)	0.950(0.004)	0.950(0.004)	0.950(0.004)	0.950(0.004)
Grade 4	0.948(0.002)	0.948(0.002)	0.948(0.002)	0.948(0.002)	0.948(0.002)
Grade 5	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.952(0.003)	0.947(0.004)
Grade 6	0.952(0.003)	0.934(0.006)	0.952(0.003)	0.952(0.003)	0.948(0.003)
Grade 7	0.954(0.003)	0.934(0.006)	0.954(0.003)	0.954(0.003)	0.950(0.003)
Grade 8	0.961(0.002)	0.938(0.007)	0.961(0.002)	0.961(0.002)	0.958(0.003)

#### 4.3.4 Summary

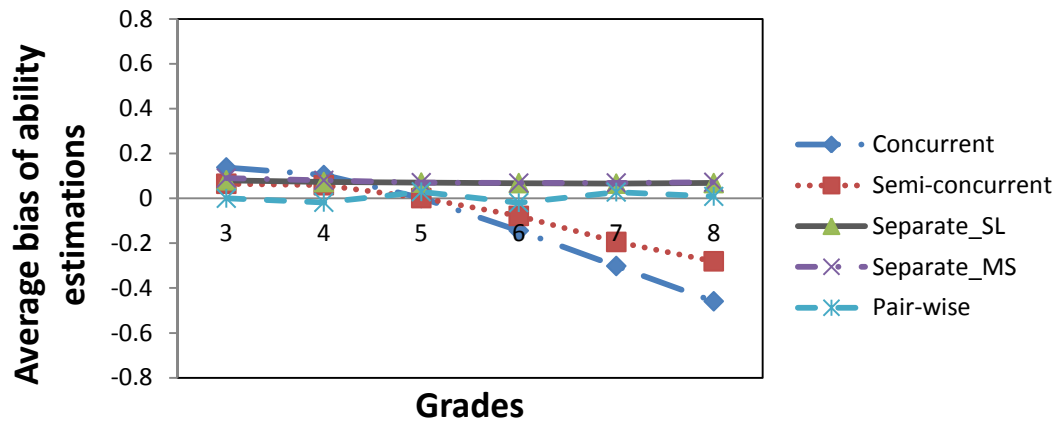
Similar to the unidimensional and multidimensional condition with  $r=0.3$ , the different variance structures of ability distributions had a small effect on the estimates of latent abilities under the multidimensional condition with  $r=0.6$  as well.

Under this multidimensional condition with  $r=0.6$ , the performance of different calibration and scaling methods were also different from those under the unidimensional condition, *i.e.*, most biases and RMSDs were much larger for most calibrations and scaling methods. Their performances were closer to those under the multidimensional condition with  $r=0.3$ , but there were differences across grades.

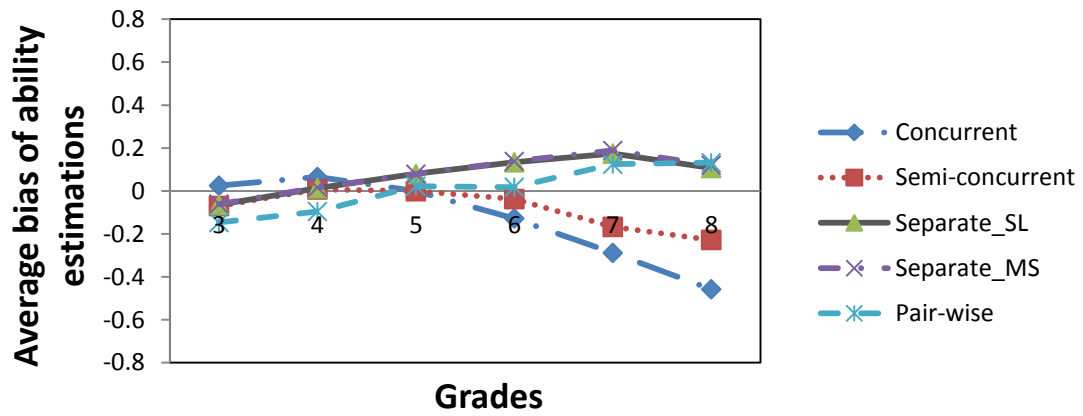
For the separate calibration with SL and MS linking methods and pair-wise calibration, the RMSDs increased consistently across most grades, with the separate calibration with MS linking increased quickly for the higher grades. For the concurrent calibration and semi-concurrent calibration, the RMSDs for the middle grades increased more than those for the grades at the two ends.

Under the multidimensional condition ( $r=0.6$ ), the average correlations between ability estimates and “true” abilities were lower than those under unidimensional condition, while higher than those the multidimensional condition with  $r=0.3$ .

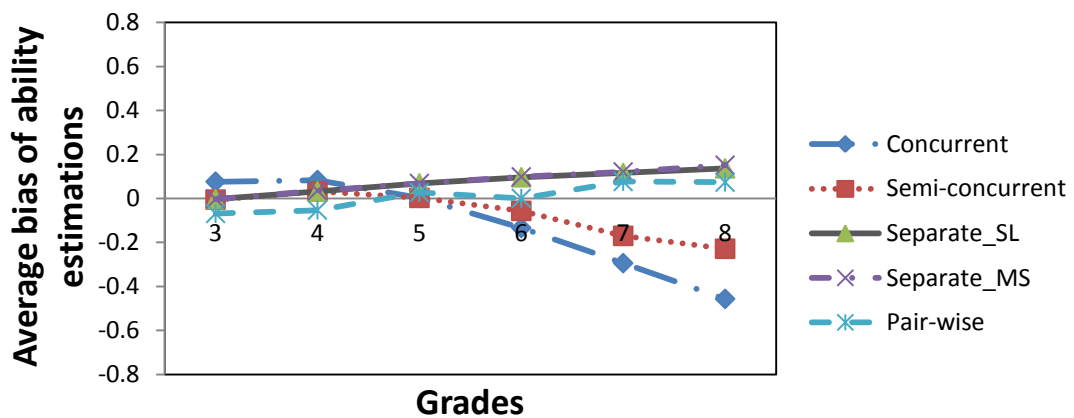
The average biases and RMSDs for the three data structures with fixed variance were compared in Figure 4.133 to Figure 4.188. The corresponding figures under varied variances conditions were not shown here because the patterns are similar to those under the fixed variance condition.



**Figure 4.13** Average bias of  $\hat{\theta}$  under unidimensional condition with fixed variance



**Figure 4.14** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance



**Figure 4.15** Average bias of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance

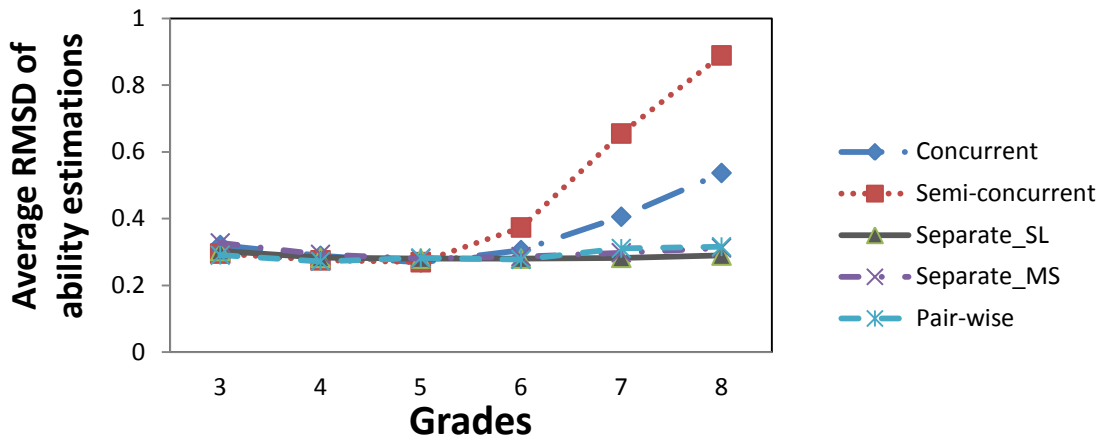


Figure 4.16 Average RMSDs of  $\hat{\theta}$  under unidimensional condition with fixed variance

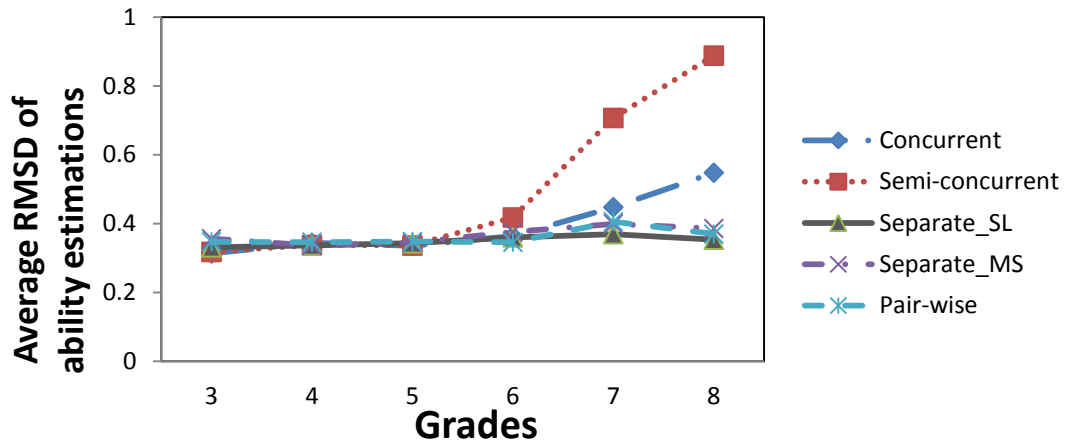


Figure 4.17 Average RMSDs of  $\hat{\theta}$  under multidimensional condition ( $r=0.3$ ) with fixed variance

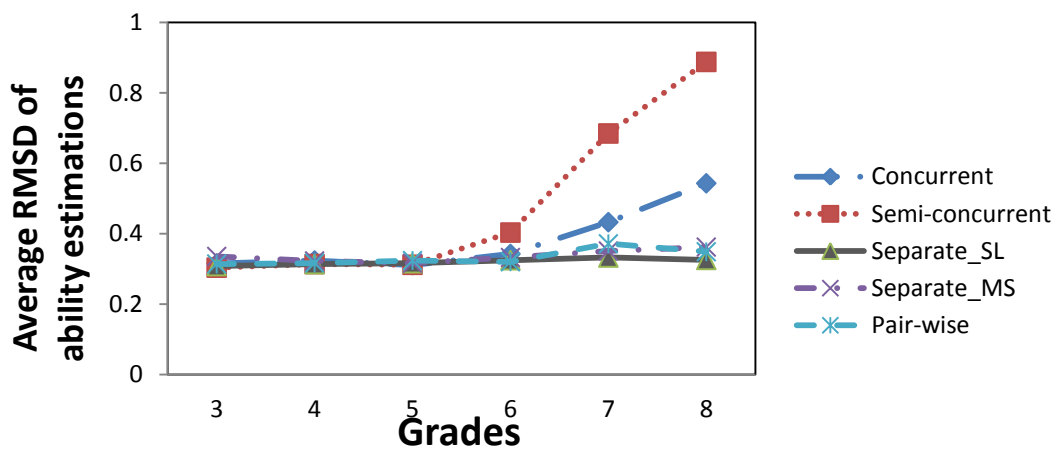


Figure 4.18 Average RMSDs of  $\hat{\theta}$  under multidimensional condition ( $r=0.6$ ) with fixed variance

In summary, under the multidimensional condition with  $r=0.6$ , ability estimation was more biased than under the unidimensional condition and has a bias pattern that is similar to that under the multidimensional conditions with  $r=0.3$ . However, the estimates were less biased for all five calibration and linking methods than those under the multidimensional conditions with  $r=0.3$ . The separate calibration with SL linking had least biased estimates across all six grades.

Furthermore, the simulated multidimensional data structures were close to essentially unidimensional data structure so as to match a real test situation, *i.e.*, the item discrimination parameter of the minor ability dimension is relative small compared to the value of the dominant ability.

In practice, the test developers try to construct the tests across grades to be approximately unidimensional or essentially unidimensional under the vertical scaling context. It is unknown to what extent the different multidimensional conditions are realistic and what degree of multidimensionality is acceptable or can be viewed as essentially unidimensional in practice. Therefore, additional studies are needed to examine real test data structures across grades.



## **5.0 DISCUSSION**

### **5.1 MAJOR FINDINGS OF THIS STUDY**

This study investigated the behaviors of different calibration and scaling methods under the common-item non-equivalent groups design. More specifically, it investigated the robustness of various unidimensional calibration and scaling procedures to the violation of the unidimensional assumption while developing vertical scales. Using common-items nonequivalent groups as the vertical scaling design, the different calibration and scaling methods were compared under both unidimensionality and multidimensionality conditions with different ability variance structures.

Three main factors were manipulated in this study. They were: 1) dimensionality, *i.e.*, unidimensional condition, multidimensional condition with  $r=0.3$ , and multidimensional condition with  $r=0.6$  (the correlations were between the dominant abilities and secondary abilities). 2) calibration and scaling methods, *i.e.*, concurrent calibration, semi-concurrent calibration, separate calibration with SL linking, separate calibration with MS linking, and pair-wise calibration. 3) variances of latent ability distributions for each grade, *i.e.*, fixed variance of 1 and varied variances across the six grades. Three measures, bias, RMSD, and correlation between estimated and “true” abilities, were examined to evaluate these calibration and linking methods.

### 5.1.1 Answers to research questions

Some of the findings in this study are consistent with those from previous studies related to calibration and linking methods. Because this study mainly focused on the robustness of unidimensional IRT-based calibration and linking methods to the violation of unidimensional assumption under the vertical scaling context, there are some additional new findings. These findings not only answer several questions with regard to the adequacy of various vertical scaling methods, but also provide practical guidance for selecting an effective vertical scaling method under different test conditions for practitioners and testing professionals.

**Question 1:** *When the IRT unidimensionality assumption holds, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma, and separate with SL(or TCC) linking, affect the vertical scaling results with assumed different latent ability variances for a wide range of grades?*

For the fixed ability variance condition, different calibration and linking methods had different effects on vertical scaling results for a wide range of grades. The separate calibration with SL and MS linking yielded the least biased ability estimates across almost all grades. The overall average bias for each grade was very small and close to 0. The average RMSDs were the smallest across a wide range of grades. The correlations between ability estimates and “true” abilities were one of the highest, which indicates that the differences between the estimated abilities and the “true” abilities were small and very consistent across the all six grades. The pair-wise calibration yielded similar results as the separate calibration with SL linking for the middle grades and slightly more biased ability estimates for the grades furthest away from the base grade.

The semi-concurrent calibration had similar ability estimates as the above three calibration methods for grades 3 to 5. However, vertical scaling results for the higher grades 6 to 8 were biased. The possible reason is that the ability transformation was used to put the ability estimates for grades 6 to 8 on the scale of grade 3 to 5. This implies that the ability transformation method is not as useful in a vertical scaling context. The vertical scaling results under the concurrent calibration were comparable to the separate calibration with SL linking methods for the base grade 5 and the grades adjacent to it. When the grades were furthest away from the base grade, however, the results were more biased, with the correlations still being very high. The most likely reason for this is that the MAP Bayesian estimation method used for ability estimation pulled the ability estimates toward the mean. Therefore, the abilities were generally overestimated for the lower grades while they were underestimated for the higher grades.

For the varied ability variance condition, the vertical scaling results under all the five calibration and linking methods were very similar to the corresponding results with fixed variance condition.

**Question 2:** *When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is low, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma, and separate with SL (or TCC) linking, affect the vertical scaling results with assumed different latent ability variances for a wide range of grades?*

When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is low, the vertical scaling results were more biased for all the five calibration and scaling methods with either

fixed variance or varied variance conditions compared to the corresponding results under the unidimensional condition. The correlations between ability estimates and “true” abilities decreased slightly for all grades.

For the fixed ability variance condition, the separate calibration with SL linking distorted the vertical scaling results consistently across grades. However, it still yielded the least biased and most consistent ability estimates across all the six grades compared among all methods. The separate calibration with MS linking and pair-wise calibration yielded slightly more biased ability estimates than separate calibration with SL linking method for some grades. It indicates that the separate calibration with MS linking was slightly more sensitive to the violation of unidimensionality assumption than the separate calibration with SL linking method. This finding is consistent with previous studies that the SL linking (or TCC method) is marginally superior in estimation accuracy compared to MS linking and mean/mean method (Baker & Al-Karni, 1991; Stocking & Lord, 1983).

Similar to the separate calibration with SL linking method, the vertical scaling results under the semi-concurrent calibration were more biased for the lower grades when compared to the unidimensional condition. However, the average deviations increased slower for the higher grades, 6 to 8, but it still yielded the most biased vertical scaling results among the five methods. For the concurrent calibration, the deviations increased more for the middle grades, 4 to 6, than the grades furthest away from the base grade, 7 and 8. The performance of the concurrent calibration was similar to separate calibration with SL linking except for at grades 7 and 8.

For the varied ability variance condition, the vertical scaling results under all the five calibration and linking methods were very similar to the corresponding results with fixed variance condition.

**Question 3:** *When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is moderate, how does the use of different calibration and scaling methods, concurrent, semi-concurrent, pair-wise calibration, separate with mean/sigma, and separate with SL(or TCC) linking, affect the vertical scaling results with assumed different latent ability variances for a wide range of grades?*

When the IRT unidimensionality assumption is violated and the correlation between the dominant dimension and secondary dimension is moderate, the different calibration and scaling methods performed differently across grades. The vertical scaling results exhibited greater bias for all five calibration and scaling methods compared to the unidimensional condition and less bias compared to the multidimensional condition with a relatively low correlation between two latent abilities. The patterns of bias were similar to the patterns under the multidimensional conditions with  $r=0.3$  for the five calibration and linking methods, with some of the biases becoming smaller.

Specifically, for the fixed ability variance condition, the separate calibration with SL linking still yielded the least biased and most consistent ability estimates across all the six grades. The separate calibration with MS linking and pair-wise calibration yielded similar ability estimates to the separate calibration with SL linking method for most grades. The concurrent calibration yielded comparable vertical scaling results as separate calibration with SL linking methods for the most grades except for grade 7 and

grade 8. The semi-concurrent calibration yielded most biased results for the higher grades, 6 to 8.

For the varied ability variance condition, the vertical scaling results under all the five calibration and linking methods were very similar to the corresponding results with fixed variance condition.

### **5.1.2 Summary of major findings**

As discussed above, the different calibration and linking methods performed differently under different conditions within the vertical scaling context. The major findings of this study can thus be summarized as follows:

**1) Difference in variance structures for ability distributions had a small effect on ability estimates and scaling results.**

To examine if the variances of latent ability distributions for the six grades had an effect on the vertical scaling results, the variances of ability distributions were either fixed at 1 for all 6 grades or varied for the 6 grades in the study. The results showed that the average biases, average RMSDs, and average correlations were close to each other under the two ability variances structures for both the unidimensional condition and the multidimensional conditions. The differences between the two variance structures were small. One possible reason for the small effect on ability estimates is that the differences among the variances of the 6 grades under the different variance structures were not very large.

**2) All methods demonstrated most consistent and least biased ability estimates for the base grade.**

Under all conditions, the deviations of ability estimates at the base grade 5 were the smallest across the six grades. The results from all five methods generally deteriorated when the linking was conducted further from the base grade. Whether the unidimensionality assumption holds or not, the standard average bias pattern for the concurrent calibration exhibited a flat “U” shape. This finding is consistent with previous research (Smith, *et. al*, 2008; Yin & Stone, 2009). For the separate calibration with linking methods and pair-wise calibration, the bias patterns were not as obvious as the concurrent calibration. However, the standard biases were still larger when the linking was conducted further from the base grade. Therefore, a middle grade should be preferred to other grades as a base grade, for it breaks the long chain of cumulated errors into two shorter halves.

**3) The various calibration and linking methods behaved differently under both unidimensional and multidimensional conditions.**

Under the unidimensional condition, the separate calibration with linking methods performed better than the concurrent method and other calibration methods for most grades. This finding is somewhat different from previous findings in Kim and Kolen’s study (2006). In their study, they found the concurrent calibration outperformed the separate calibration with different linking methods in linking accuracy and robustness to format effects (the use of mixed-format tests such as both multiple-choice items and constructed-response items). But the performance of concurrent calibration was only slightly better than that of characteristic curve methods such as SL and Haebara linking methods. The possible reason for this discrepancy is that only two level

groups were used in their linking design, while a much wider range of groups were scaled in this vertical scaling context. In addition, in the current study, MAP, a Bayesian method, was used for ability estimation with MULTILOG. The Bayesian method may lead to the shrinkage of the whole scale with a wide range of grade levels.

However, this finding is partially consistent with previous studies (Karkee, *et.al*, 2003; Lee & Ban, 2010). In Lee and Ban' study, they found that the separate estimation method produced consistently better results than did the concurrent calibration under a random groups equating design. Karkee and his colleagues had similar findings in a vertical scaling study with real data. In current study, for the base grade 5 and the two adjacent grades, the concurrent calibration yielded very similar results to separate calibration, but its standard average biases increased more quickly for grades 3, 7 and 8. On the other hand, the correlations between ability estimates and "true" abilities under concurrent calibration and separate calibration were almost identical. These results indicate that the rank of students' abilities did not change much under concurrent calibration, but the mean and variances of the ability distributions differed more when the grades were further from the base grade 5. It implies that the concurrent performs relatively well for a narrow range of grades.

Separate calibration with SL linking provided slightly lower average RMSDs and standar deviations of RMSDs at very low and very high grades and under the multidimensional condition when compared to separate calibration with MS linking. This means the separate calibration with SL linking performed slightly better and provides slightly more stable results than the separate calibration with MS linking when the grades to be linked were further away from the base grade or when the unidimensionality assumption was violated. It also implies that separate calibration with



SL linking is somewhat less sensitive to atypical groups and items. This finding is also consistent with previous studies that the SL linking (or TCC method) is marginally superior in estimation accuracy compared to MS linking and mean/mean method (Baker & Al-Karni, 1991; Stocking & Lord, 1983). As discussed in Chapter 2, the MS linking method is based on the statistics of item difficulties and item discriminations. At the extreme grades, the item difficulties could be more atypical than the items for the middle grades. Therefore, the MS linking is more sensitive to atypical groups and items in the extreme grades.

The performances of the two hybrid calibrations used in this study also differed from each other. The pair-wise calibration provided very similar estimates as the separate calibrations. Its performance is similar to the separate calibration with MS linking and comparable to the separate calibration with SL linking. Furthermore, the pair-wise calibration performed better than the concurrent calibration for the two extreme grades. This finding is somewhat expected since the hybrid calibrations are expected to strengthen the ability estimates of concurrent calibration and separate calibration methods.

The other hybrid calibration method, the semi-concurrent calibration, provided comparable results for grades 3 to 5 as the separate calibration with SL linking and the pair-wise calibration. It also performed slightly better than the concurrent calibration. However, for grades 6 to 8, the results were more biased when compared to the other methods. The possible source for this discrepancy is the difference in the linking methods used under the two hybrid calibrations. In pair-wise calibration, the SL linking was used to link each pair of calibrations, while the proficiency transformation was used to link two separate concurrent calibrations for two sets of grades (grades 3, 4, 5 and

grades 5, 6, 7, 8) in the semi-concurrent calibration. The results from the pair-wise calibration with SL linking further support the use of SL linking in the vertical scaling context. Therefore, the proficiency transformation may not be an acceptable transformation method in vertical scaling with a wide range of grades.

**4) Multidimensionality in item responses clearly affects the performance of unidimensional IRT-based calibration and linking methods in vertical scaling, particularly across a wide range of grades.**

Under the multidimensional data structure, all ability estimates were more biased than under the unidimensional condition for all five calibration and linking methods. In addition, the correlation between the ability estimates and “true” abilities decreased.

First, the degree of relationship between the dominant ability and secondary ability had an effect on the ability estimates. When ability on the secondary dimension had a low correlation (such as 0.3) with the ability on the dominant dimension, the ability estimates on the dominant ability were biased when unidimensional vertical scaling methods were used. The correlations decreased quickly, especially for the middle grades, 4 to 7. When ability on the secondary dimension was moderately related to the ability on the dominant dimension, the ability estimates were less biased than the above multidimensional data structure. This implies that when ability on the secondary dimension has a higher correlation with the ability on the dominant dimension, the vertical scaling results would be more robust to the violation of the unidimensionality assumption.

Second, the influences of multidimensionality exhibit different patterns for different grades with different calibration and linking methods. For the separate calibration with SL linking and pair-wise calibration, the multidimensional data structure

had a consistent effect on all grades. For the separate calibration with MS linking, the multidimensional data structure had a slightly larger effect when the grades were further away from the base grade. For the concurrent calibration, the multidimensional data structure had a larger effect for the middle grades than for the grades at the two ends. For semi-concurrent calibration, the multidimensional data structure also had a larger effect on the lower grades than the higher grades.

Some of these findings are consistent with findings from previous studies. One study (Beguin, *et. al*, 2000) showed that multidimensionality of data affects the relative performance of separate and concurrent unidimensional estimation methods under the horizontal equating context. It was found that unidimensional IRT models resulted in reasonable estimates when applied to multidimensional data from an equivalent groups design, but unidimensional IRT models led to large deviations between true and estimated score distributions when using multidimensional data from a non-equivalent groups design under the horizontal equating context.

In the current vertical scaling study, multidimensionality also affected the performance of calibration and linking methods on ability estimates. When unidimensionality assumption holds, the difference between ability estimates and “true” abilities were small and correlations were high for most grades, except for the concurrent calibration and semi-concurrent calibration for the two higher grades 7 and 8. When unidimensionality is violated, the deviations between ability estimates and “true” abilities generally were higher, while the correlations were lower than the corresponding results under the unidimensional condition. This indicates that in addition to increases in the standard biases, the rank of students’ abilities changed under the multidimensional data structure.

### **5.1.3 Implications to educators and research practitioners**

The findings discussed above have several implications to educators and practitioners. First, the selection of an effective vertical scaling method should be based on an examination of the item response structure. If the item response structure is unidimensional or essentially unidimensional, the separate calibration with SL and MS linking should be considered for conducting vertical scaling. The pair-wise calibration should also be considered, because it provides comparable ability estimates as separate calibration with SL and MS linking. Semi-concurrent calibration with SL linking may also be an option. However, the proficiency transformation is not good method for vertical scaling. If the unidimensionality assumption is highly violated, the separate calibration with SL linking is the best option for conducting vertical scaling.

Second, a middle grade is preferred to other grades as a base grade. Selecting a middle grade as the base grade breaks the measurement errors from a long chain into two shorter parts. It reduces errors for those grades which are further away from the base grade, especially for certain methods such as concurrent calibration.

Third, narrowing the range of scaled grade levels could be considered when there is a violation of unidimensionality across a wide range of grades. As discussed in Chapter 2, the unidimensionality assumption in vertical scaling tests refers to unidimensionality in each test and construct invariance across grades, *i.e.*, the tests should measure the same or very similar construct and focus on only a single latent trait across wide grades (Li & Lissitz, 2012). However, if it is known that there is a significant construct shift across a wide range of grades, constructing a vertical scale within a narrow range of grades may be a better alternative than within a very wide

range of grades. Based on the results of the current study, the vertical scaling results were less biased for the middle grades than for either the lower grades or the higher grades. If the vertical scaling is conducted with a narrower range, the MS linking method and proficiency transformations are not recommended based on the results of this study. Separate calibration with SL linking method, hybrid calibration with SL linking method, and concurrent calibration should be considered for vertical scaling.

## 5.2 LIMITATIONS AND FUTURE STUDY

This simulation study compared the performance of various IRT-based unidimensional calibration and linking methods under different test conditions in the vertical scaling context. Vertical scaling is a complicated process because a lot of technical and practical issues influence the scaling results. Therefore, the conclusions and recommendations should be considered in light of the limitations of this simulation study.

First, the different test conditions are based on literature reviews. It is unknown to what extent the different multidimensional conditions are realistic. Therefore, additional studies are needed to examine real test data structures across grades.

Second, in this study, the number of common items (30% of the number of total items) and total number of items were fixed for all test conditions. In practice, the number of common items may not always be the same across wide grades. In future study, different levels of the number of common items such as 20% or 40% of total items could be manipulated under different test conditions.

Furthermore, in this simulation study, the common items were randomly selected from lower level grade items. The random selection may not represent the real situation in practice. Ideally, the content and difficulty of common items should span multiple grades. If these common items are on-level items for one grade, they could be out-of-level items for the adjacent grades, either easier or more difficult for the adjacent grade students. If students are given items that are extremely too difficult or easy for them, the resulting data are likely of poor quality with associated “floor” and “ceiling” effects. Using tests of inappropriate difficulty typically also leads to large conditional standard errors of measurement (CSEM) which reduces the reliability of the test scores. More

importantly, the associated measurement errors are considered as another artificial cause of scale shrinkage. In future study, the difficulties of each common item blocks could be constrained in a specified range.

Finally, only dichotomously scored items were used in this study. The multidimensional item response structures were simulated to be close to essential unidimensional structure. However, in practice, tests are usually constructed with both dichotomously scored and polytomously scored items. For the mixed format tests, the multidimensional data structure could be more complex. Future study can be extended to mixed format tests to examine the behaviors of different calibration and linking methods under the vertical scaling context.

## APPENDIX A

### SELECTED MPLUS RESULTS

**Table 1: Geomin rotated factor loadings**

Items	Factor loadings on factor1	Factor loadings on factor2
Item1	0.592	0.284
Item 2	0.532	0.258
Item 3	0.614	0.188
Item 4	0.527	0.216
Item 5	0.363	0.324
Item 6	0.612	0.204
Item 7	0.463	0.19
Item 8	0.709	0.206
Item 9	0.639	0.241
Item 10	0.617	0.251
Item11	0.79	-0.039
Item 12	0.774	-0.068
Item 13	0.843	-0.051
Item 14	0.49	-0.078
Item 15	0.557	0.116
Item 16	0.519	-0.093
Item 17	0.776	0.077
Item 18	0.585	0.006
Item 19	0.38	0.048



Item 20	0.415	0.016
Item21	0.744	-0.098
Item 22	0.405	0.015
Item 23	0.731	0.002
Item 24	0.642	-0.025
Item 25	0.713	0.036
Item 26	0.516	0.084
Item 27	0.46	0.082
Item 28	0.778	-0.036
Item 29	0.697	-0.032
Item 30	0.489	0.034
Item31	0.606	-0.144
Item 32	0.682	-0.048
Item 33	0.542	-0.03
Item 34	0.571	0.144
Item 35	0.617	0.006
Item 36	0.344	0.08
Item 37	0.746	-0.131
Item 38	0.541	0.078
Item 39	0.783	-0.022
Item 40	0.215	0.2
Item 41	0.585	-0.028
Item 42	0.452	0.099
Item 43	0.519	-0.108
Item 44	0.731	0.148
Item 45	0.509	0.023
Item 46	0.575	0.053
Item 47	0.367	0.111
Item 48	0.512	-0.009
Item 49	0.471	-0.058
Item 50	0.555	-0.044

**Table 2: Factor structure**

Items	Factor1	Factor2
Item1	0.658	0.422
Item 2	0.592	0.382
Item 3	0.658	0.331
Item 4	0.577	0.339
Item 5	0.438	0.409
Item 6	0.659	0.347
Item 7	0.507	0.298
Item 8	0.757	0.371
Item 9	0.695	0.389
Item 10	0.675	0.394
Item11	0.781	0.145
Item 12	0.759	0.112
Item 13	0.831	0.146
Item 14	0.471	0.036
Item 15	0.584	0.245
Item 16	0.497	0.028
Item 17	0.794	0.258
Item 18	0.587	0.143
Item 19	0.391	0.137
Item 20	0.419	0.113
Item21	0.721	0.075
Item 22	0.408	0.109
Item 23	0.731	0.173
Item 24	0.636	0.125
Item 25	0.721	0.202
Item 26	0.535	0.205
Item 27	0.48	0.19
Item 28	0.77	0.146
Item 29	0.69	0.131
Item 30	0.497	0.148

Item31	0.572	-0.003
Item 32	0.671	0.111
Item 33	0.535	0.096
Item 34	0.605	0.278
Item 35	0.618	0.15
Item 36	0.363	0.161
Item 37	0.715	0.043
Item 38	0.559	0.204
Item 39	0.778	0.161
Item 40	0.261	0.25
Item 41	0.578	0.108
Item 42	0.476	0.205
Item 43	0.493	0.013
Item 44	0.765	0.318
Item 45	0.515	0.142
Item 46	0.587	0.187
Item 47	0.393	0.196
Item 48	0.51	0.11
Item 49	0.457	0.052
Item 50	0.545	0.085

## APPENDIX B

### SAS PROGRAM TO SIMULATE VERTICAL SCALING

#### B.1 GENERAL MACRO

```
libname liquin 'd:\dissertation';
```

```
%macro loop_dissertation;  
%do toprep=1 %to 200;  
X 'cd d:\dissertation\multilog';  
x 'exit';  
%include 'd:\dissertation\item_responses_15c_0118.sas';
```

```
/*macro for concurrent calibration*/
```

```
x 'mlg response_38_PAR';  
x 'mlg response_38_map';
```

```
%include 'd:\dissertation\rmsd_concurrent.sas';  
proc append BASE=liquin.final_rmsd_con data=liquin.rmsd_con;
```

```
/*macro for semi-concurrent calibration*/
```

```
X 'cd d:\dissertation\multilog';  
x 'exit';
```

```
%include 'd:\dissertation\semi_con.sas';
```

```
x 'mlg response_35_PAR';
```

```

x 'mlg response_35_MAP';
x 'mlg response_58_PAR';
x 'mlg response_58_MAP';

%include 'd:\dissertation\rmsd_semi.sas';

proc append BASE=liquun.final_rmsd_semi data=liquun.rmsd_semi;

/*macro for separate calibration*/

%include 'd:\dissertation\separate_grade38.sas';

X'cd d:\dissertation\multilog';
X 'mlg response_3_PAR';
x 'mlg response_3_MAP';

X 'mlg response_4_PAR';
x 'mlg response_4_MAP';

X 'mlg response_5_PAR';
x 'mlg response_5_MAP';

X 'mlg response_6_PAR';
x 'mlg response_6_MAP';

X 'mlg response_7_PAR';
x 'mlg response_7_MAP';

X 'mlg response_8_PAR';
x 'mlg response_8_MAP';
X 'exit';

%include 'd:\dissertation\separate_ST_Linking.sas';

x 'cd d:\dissertation\ST';
x 'st_wc_cmd'
x 'exit';

%include 'd:\dissertation\rmsd_separate.sas';

proc append BASE=liquun.final_rmsd_sep data=liquun.rmsd_sep;

/* macro for pairwise*/

```

```

%include 'd:\dissertation\pairwise_grade38.sas';

x 'cd d:\dissertation\multilog';

X 'mlg response_34_PAR';
x 'mlg response_34_MAP';

X 'mlg response_56_PAR';
x 'mlg response_56_MAP';

X 'mlg response_78_PAR';
x 'mlg response_78_MAP';

X 'exit';

%include 'd:\dissertation\pairwise_ST_Linking.sas';

x 'cd d:\dissertation\ST';
x 'st_wc_cmd_pair'
x 'exit';

%include 'd:\dissertation\rmsd_pairwise.sas';

proc append BASE=liquun.final_rmsd_pair data=liquun.rmsd_pair;

data liquun.rmsd_all_r06;

merge liquun.final_rmsd_con liquun.final_rmsd_semi liquun.final_rmsd_sep
liquun.final_rmsd_pair;
run;

%end;
%mend;
%loop_dissertation
run;

```

## B.2 ITEM RESPONSE SIMULATION

```
libname liqun 'd:\dissertation';

%let nsize=2000;

/*generating a dataset under 3_p model with D=1.702*/

proc iml;
seed = 0;
n = &nsize;
nitem = 50;
overlap = 15;
g=6;
r1=0;
r2=0;
nGroup=6;
d=0;
u = 0.0;
x = 9;
fseed = 0;
lseed = 0;
rho=0.6;
yparam = {-1, -0.45, 0, 0.35, 0.6, 0.8};
dparam = {-1, -0.45, 0, 0.35, 0.6, 0.8};

param=J((nitem + (ngroup-1)*(nitem-overlap)), 3, 0);

file 'd:\dissertation\param_1g.dat';
k=1;

do g=1 to nGroup;

if g=1 then do
j= 1 to nitem;
a = 0.75*ranuni(seed)+ 0.75;
d = 0.75*rannor(seed) -dparam[g,1];
c = 0.2;
param[k,1] = a;
param[k,2] = d;
param[k,3] = c;
put a +1 d +1 g +1 c;
```

```

        k=k+1;
end;

if g>=2 then do
    j= 1 to (nitem-overlap);
        a = 0.75*ranuni(seed)+ 0.75;
        d = 0.75*rannor(seed)-dparam[g,1];
        c = 0.2;
        param[k,1] = a;
        param[k,2] = d;
        param[k,3] = c;
    put a +1 d +1 g +1 c ;
        k=k+1;
    end;
end;

file 'd:\dissertation\MIRT_15c.dat';

do g=1 to nGroup;

    do j= 1 to n;
        call rannor(seed, r1);
        call rannor(seed, r2);
        y1=r1+yparam[g,1];
        y2=rho*r1+sqrt(1-rho**2)*r2+yparam[g,1];
        put j 4.0 +1 y1 6.3 +1 y2 6.3 +1 g +1@;

do m =1 to nGroup;
    if m=g then do;
        base= (m-1)*(nitem -overlap);
do i=1 to nitem;

        a = param[base+i,1];
        d = param[base+i,2];
        c = param[base+i,3];

        if ((i<=overlap & g>1) | (i >(nitem-overlap) & g<6)) then
            z = 1.702*(a*y1+0.4*y2+d);
        else

            z = 1.702*(a*y1+d);
            u =ranuni( seed );
            p=c+(1-c)*exp(z)/(1+exp(z));
            if u < p then
                resp=1;
            else resp=0;

```



```

put resp 1.0 @;
end;

end;

else do
i=1 to (nitem-overlap);
put x 1.0 @;
end;

end;

put;

end;

end;

quit;
run;

```

## **APPENDIX C**

### **SELECTED MULTILOG FILES**

#### **C.1 MULTILOG FILES FOR CONCURRENT CALIBRATION**

##### **C.1.1 Getting Item Parameter Estimations**

```
>PROBLEM RANDOM,  
  
INDIVIDUAL,  
  
DATA = 'd:\dissertation\MIRT_15c.dat',  
  
NITEMS = 225,  
  
NGROUPS = 6,  
  
NEXAMINEES = 12000,  
  
NCHARS =16;  
  
>TEST ALL,  
  
    L3;  
  
>SAVE;  
  
>END ;  
  
3  
  
019
```



[illegible]

## C.2 MULTILOG FILES FOR SEPARATE CALIBRATION

### C.2.1 Getting Item Parameter Estimations

>PROBLEM RANDOM,

INDIVIDUAL,

DATA = 'd:\dissertation\g3\_response.dat',

NITEMS = 50,

NGROUPS = 1,

NEXAMINEES = 2000,

NCHARS =17;

>TEST ALL,

L3;

```
>SAVE;
```

```
>END ;
```

3

019

111

Y

9

(5a1,1x,6a1,1X,6a1,1x,50a1)

[illegible]

## BIBLIOGRAPHY

- Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7, 255-278.
- Bejar, I. I. Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, 6, 309-325.
- Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, 8(3), 261-171.
- Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement*, 20, 45-57.
- Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.
- Béguin, A. A., Hanson, B. A., & Glas, C.A.W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. Available from <http://www.ba-h.com/papers/paper0002.html>
- Boughton, K. A., Lorié, W., & Yao, L. (2005). A Multidimensional Multi-Group IRT Model for Vertical Scales with Complex Test Structure: An Empirical Evaluation of Student Growth using Real Data. *Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada*.
- Briggs, D. (2008). An introduction to multidimensional IRT. Paper presented at UC Berkeley (April).
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational and Measurement: Issues and Practice*, 28 (4), 3-14.
- Briggs, D. C., Weeks, J. P., & Wiley, E. (2008). Vertical Scaling in Value-Added models for Student Learning.

[http://www.colorado.edu/education/faculty/derekbriggs/Docs/BWW\\_VSVAM\\_Wisconsin\\_040708.pdf](http://www.colorado.edu/education/faculty/derekbriggs/Docs/BWW_VSVAM_Wisconsin_040708.pdf)

- Camilli, G., Yamamoto, K., Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, 17(4), 379-388.
- Chin, T.-Y., Kim, W., & Nering, M. L. (2006). *Five statistical factors that influence IRT vertical scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment*, 1(4), 329-347.
- Cook, L.L. & Eignor, D.R. (1991). NCME Instructional module: IRT equating methods. *Educational Measurement: Issues and Practice*, 10, 37-45.
- Cook, L.L. & Eignor, D.R., & Taft, h. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.
- CSAP. (2007). Colorado State Assessment Program Technical report.
- De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, 33(2), 181-201.
- DeMars, C. E. (2003). Detecting multidimensionality due to curricular differences. *Journal of Educational Measurement*, 40, 29-51.
- DeMars, C. E. (2003). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43, 145-168.
- Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, 9, 413-415.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fitzpatrick, A. R. & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, 14(1), 31-57.
- FCAT. (2001). Florida Comprehensive Assessment Test Technical Report.
- FCAT. (2006). Florida Comprehensive Assessment Test Technical Report. <http://fcat.fldoe.org/fcatpub5.asp>



- Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied psychological measurement, 35*(1), 67-82.
- Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.
- Haertel, E. H. (2005). Using a longitudinal student tracking system to improve the design for public school accountability in California. <http://ed.stanford.edu/suse/faculty/haertel/Haertel-Value-Added.pdf>
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.
- Hanson, B. & Zeng, L. (2004). ST: ST. (2004). A computer program for IRT scale transformation. *Center for Advanced Studies in Measurement and Assessment*, University of Iowa.
- Harris, D. J. (2007). Practical issues in vertical scaling. In Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.), *Linking and aligning scores and scales*. New York: Springer Science + Business media, Inc.
- Harris, D. J. & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151-159.
- Harwell, M.R., Stone, C.A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.
- Holmes, S. E. (1982). Unidimensionality and Vertical Equating with the Rasch Model. *Journal of Educational Measurement, 19*(2):139-147.
- Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32* (4), 311-333.
- Ito, K., Sykes, R. C., & Yao, L (2008). Concurrent and Separate Grade-Groups Linking Procedures for Vertical Scaling. *Applied Measurement in Education, 21*, 187-206.
- Karkee, T., Lewis, D. M, Hoskens, M., Yao, L., & Haug, C. A. (2003, April). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the Annual Meeting of the National Conference on Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED478167).
- Kaskowitz, G.S. & De Ayala, R.J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*, 39-52.

- Kim, S.-H. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22 (2), 131-143.
- Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26, 25-41.
- Kim, J., Lee, W., Kim, D., Kelley, K. (2009). Investigation of vertical scaling using the Rasch model. *Paper was presented at the annual meeting of the National council on measurement in Education, San Diego, CA.*
- Kim, J. K. & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*. 58 (4), 587-599.
- Kim, S. & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, 19(4), 357-381.
- Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8 (2), 147-154.
- Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking* (2nd). NY: Springer Science + Business media, Inc.
- Lee, W. -C. & Ban J.-C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, 23, 23-48.
- Lee, W., Song, M., & Kim, J. (2004). An Investigation of Procedures for obtaining a Common IRT Scale. *Paper presented at the Annual Meeting of the American Education Research Association, San Diego, April 2004.*
- Li, Y. & Lissitz, R.W. (2012). Exploring the Full-Information Bifactor Model in Vertical Scaling With Construct Shift. *Applied Psychological Measurement*, 36, 3-20.
- Linn, R. L. (2000). Assessments and accountability. *Educational researcher*. 29 (2), 4-16.
- Linn, R. L., Levine, m. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159-173.
- Lissitz, R. W. & Huynh. H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10).
- Liu, J. & Walker, M. E. (2007). Score linking issues related to test content changes. In Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.), *Linking and aligning scores and scales*. New York: Springer Science + Business media, Inc.

- Lockwood, J.R. McCaffrey, D. F., Hamilton, L.S., Stecher, B. Le, V.N, & Martinez, J. F. (2007) The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*. 44 (1).47-67.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179–193.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139–160.
- Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*. 31 (1).35-62.
- McBride, J & Wise, L (2001). Developing a Vertical Scale for the Florida Comprehensive Assessment Test, Harcourt Educational Measurement report.
- McCall, M. (2007). Vertical scaling and the development skills. Paper presented at WERA/OSPI state assessment conference, SeaTac, WA.
- McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability (MG-158-EDU)*. Santa Monica, CA: RAND.
- McLaughlin, D. (2005). Considerations in using the longitudinal school-level state assessment score database. Paper presented at the Symposium on the use of school-level data for evaluating Federal Education Programs. [http://www7.nationalacademies.org/BOTA/School-Level%20Data\\_McLaughlin-Final.pdf](http://www7.nationalacademies.org/BOTA/School-Level%20Data_McLaughlin-Final.pdf)
- Meng, H. (2007). A comparison study of IRT calibration methods for mixed-format tests in vertical scaling. Un-published dissertation in University of Iowa.
- Meyers, J. L., Miller, G. E., & Way W. D., (2009). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied Measurement in Education*, 22, 38-60.
- MULTILOG (2003). Version 7.0 <http://www.ssicentral.com/>.
- NCLB. (2001). U.S. Department of Education Website. <http://ed.gov/nclb/landing.jhtml>
- Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, 25, 373–383.

- Patz, R. J (2007). Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems. *Prepared for the technical issues in large scale assessment (TILSA)*. <http://www.ccsso.org>.
- Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.), *Linking and aligning scores and scales*. New York: Springer Science + Business media, Inc.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reckase, M. D. & Martineau, J. (2004). The vertical scaling of science achievement tests. *Paper commissioned by the committee on test design for K-12 Science achievement, center for education, national research council*.
- Rupp A. A. & Zumbo, B. D. (2006). Understanding parameter invariance in Unidimensional IRT models. *Educational and Psychological Measurement*, 66, 63-84.
- Schmidt, W. H, Houang, R. T., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-Added Models in Education: Theory and Applications* (pp. 145-164). Maple Grove, MN: JAM Press.
- Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159-194.
- Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological measurement*, 10, 303-317.
- Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, 12(1), 69-82.
- Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 15, 23-35.
- Smith, Z. R., Finkelman, M., Nering, M. L., & Kim, W. (2008). Vertical Scaling: A Comparison of Linking Methods with Unidimensional and Multidimensional Data. <http://www.measuredprogress.org/resources/psychometrics/framework/materials/08/VertScal.Smith.pdf>
- ST. (2004). A computer program for IRT scale transformation. *Center for Advanced Studies in Measurement and Assessment*, University of Iowa.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 210-210.

- Stone, C. A. (2007). Item response theory lectures. *School of education, University of Pittsburgh*.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.
- Tong, Y. & Kolen, M. J. (2007) Comparisons of Methodologies and Results in Vertical Scaling for Educational Achievement Tests. *Applied Measurement in Education*, 20 (2), 227-253.
- Walker, C.M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: the effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological measurement*, 66(5), 721-738.
- Wang, X. & Harris, D. J. (2009). Maintenance of Vertical Scales Under Common item Nonequivalent Groups Design. *Paper was presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA*.
- Wang, S., Jiao, H., Young, M., & Jin, Y. (2005). The effects of linking designs in vertical scaling on the growth patterns of student achievement.
- Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, 35, 93-107.
- Wilson, M. (2009). Growth in student achievement: Can we have both meaning and technical rigor? *Paper was presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA*.
- Yao, L., & Mao, X. (2004). Unidimensional and multidimensional estimation of vertical scaled tests with complex structure. Paper was presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299-325.
- Yen, W. M. & Burket, G. R. (1997). Comparison of Item Response Theory and Thurstone Methods of Vertical Scaling. *Journal of Educational Measurement*. 34 (4). 293-313.
- Yin, L. & Stone, C. (2009). Comparisons of IRT-Based Ability Estimation Methods and Calibration Procedures in Scaling Test Design. *Paper was presented at 40<sup>th</sup> Annual Northeastern Educational Research Association Conference, Rocky Hill, CT*.
- Zhang, B. & Stone, C. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*. 68 (2), 181-196.